A review of multiple hypothesis testing in relation to the use of lateral cephalometric variables as the outcome measure in orthodontic research

Thesis submitted in accordance with the requirements of the University of Liverpool for the Degree of Doctor of Dental Science (Orthodontics)

by

Shih Chia Pua

June 2018

Acknowledgement

I would like to take this opportunity to express my greatest gratitude to both my research supervisors Dr Burnside (GB) and Dr Flannigan (NF) for their constructive advice and guidance during the planning and execution of this research project. Without their enthusiastic encouragement and useful feedback, completion of this project would not have been possible.

I am sincerely grateful to the Malaysian Government for making my dream come true and without doubt this educational journey has been a pleasant and rewarding experience for me.

I would like to thank my parents and siblings for their spiritual support and encouragement throughout my whole study period. Thank you for having faith in me and standing by my side whenever I need them.

Not forgetting my friends, Pek Ying, Ying Yee, Fei Chi, Leng Siang and Lee Boon, for their continuous support and encouragement. Special mention to Amy, thank you very much for your time and effort during the stage of my research protocol development.

Finally, I wish to thank my husband for his endless love, support and patience by my side in this research journey.

Abstract

A review of multiple hypothesis testing in relation to the use of lateral cephalometric variables as the outcome measure in orthodontic research

Aim:
To examine the extent of the multiple hypothesis testing and its correction in orthodontic research in relation to the use of lateral cephalometric variables as the outcome measure.

Study design:
A retrospective, observational study looking at a sample of published orthodontic articles (n=1688) over a two-year period from 1st January 2014 to 31st December 2015.

Data sources:
Four major electronic databases namely PubMed, Ovid Medline, Scopus and EBSCO Dentistry & Oral Sciences Source were electronically searched using Medical Subject Heading (MeSH) terms. Additionally, all issues of American Journal of Orthodontics and Dentofacial Orthopedics (AJODO), The Angle Orthodontist (AO), European Journal of Orthodontics (EJO) and Journal of Orthodontics (JO) were also hand-searched systematically. Both searches were carried out independently by first author (SCP).

Review methods:
Eligible articles were identified and reviewed independently by first author (SCP) to determine whether the articles tested greater than five hypotheses in at least one family of inferences with respect to the predetermined criteria. For articles meeting the criterion for multiple testing, type I error rates were calculated. Additionally, a statistical correction experiment using Bonferroni's method was applied to the reported results of the included studies.

Additional information was collected on: study type (prospective/ retrospective), journal classification (main/ non-main orthodontic journal were classified based on 2015 SCImago Journal and Country Rank), region of authorship (Americas, Europe and Asia/ others), number of researcher in the publication (1-4, 5-7 and 8 or more) and involvement of a statistician to examine whether these factors were associated with multiple testing correction.

Results:
Of the 139 studies associated with multiple testing, there was approximately 3 families of tests (per article) with an average of 20 hypothesis tests (range 5-47) using lateral cephalometric

variables as the outcome measure (per family of tests). Only 40 publications (29%) considered the effect of multiple testing that these studies in some way have corrected or accounted for multiple testing.

Within the studies that have not accounted for multiple hypothesis testing, there was a mean 58% chance of committing a type I error and, on average, 13% of the significant results were likely to be false positives. After the application of the Bonferroni's method in the correction experiment, only 47% of the significant results reported within the articles that remained significant.

Studies published in the main orthodontic journals (AJODO, EJO, JO and KJO) were more likely to account for multiple testing ($p=0.002$). Handsearching was superior than electronic searching with 10% of papers ($n=5$) which were missed from electronic searching.

Conclusions:

Multiple testing is common in the orthodontic research especially in relation to the use of multiple cephalometric variables as the outcome measure. This study demonstrates that the risk of false positive findings is considerably high and only a minority of the articles that have in some way corrected or accounted for multiple testing. Therefore, this multiplicity issue in relation to the use of multiple cephalometric variables in a cephalometric study deserves a closer attention from researchers, reviewers and readers.

Table of Contents

List of abbreviations

MANOVA	Multivariate Analysis of Variance
ANOVA	Analysis of Variance
Tukey's HSD	Tukey's Honestly Significant Difference
AJODO	American Journal of Orthodontics and Dentofacial Orthopedics
χ^2	Chi-square test
NHP	Natural head position
RCP	Retruded contact position
UK	United Kingdom
SD	Standard deviation
LI	Lower central incisor
MP	Mandibular plane
FH	Frankfort horizontal plane
Na	Nasion
S	Sella
Ar	Articulare
Go	Gonion
Gn	Gnathion
PFH	Posterior facial height
AFH	Anterior facial height
VTO	Visualised treatment objective
SN line	Sella-Nasion line
H line	Harmony line
Or	Orbitale
Ba	Basion
Bo	Bolton Point
ANS	Anterior nasal spine
PNS	Posterior nasal spine
Ptm	Pterygomaxillary fissure
A	Point A (subspinale)
B	Point B (supramentale)
Pog	Pogonion
Me	Menton
Po	Porion
G	Glabella
Ils	Inferior labial sulcus

Li	Labrale inferius
Ls	Labrale superius
Ms	Menton soft tissue
Ns	Nasion soft tissue
Pn	Pronasale
Pos	Pogonion soft tissue
Sls	Superior labial sulcus
Sn	Subnasale
St	Stomion
Sti	Stomion inferius
Sts	Stomion superius
RCT	Randomised clinical trial/ Randomised controlled trial
CBCT	Cone beam computed tomography
MeSH	Medical subject headings
AO	The Angle Orthodontist
EJO	European Journal of Orthodontics
JO	Journal of Orthodontics
SJR	SCImago Journal and Country Rank
IQR	Interquartile range
KJO	Korean Journal of Orthodontics
IF	Impact factor
JCR	Journal of citation report
ARVO	The Association for Research in Vision and Ophthalmology
OPO	Ophthalmic & Physiological Optics
OVS	Optometry & Vision Sciences
CXO	Clinical & Experimental Optometry
FOV	Field of view
CIs	Confidence intervals

List of tables

List of figures

Chapter 1: Introduction

Hypothesis testing is the process of deciding statistically whether the findings of an investigation reflect real associations or chance at a given level of probability.[1] The purpose of hypothesis testing is to assist the researchers in reaching a conclusion about the population by drawing inferences from a sample of the studied population.[2] The results are then translated into p values and are used to define whether the results of the test are either significant or non-significant.[1,3] The threshold value for the level of significance, denoted by alpha (α), is arbitrary and usually set in advance. Conventionally, most studies set an alpha (α) level of 0.05.[1,3,4]

The p value represents the probability of obtaining the observed difference by chance, if the null hypothesis is true.[1,2] A p value of less than the threshold value for significance is described as statistically significant. Therefore, one can conclude that an observed significant difference is unlikely to have occurred by chance alone.[1] In other words, it is also the probability of rejecting null hypothesis [H_0] when it is true and thus accepting the alternative hypothesis [HA]. This is known as a type I error.[1–7]

Multiple hypothesis testing refers to carrying out a number of significance tests on a data set within a study.[1,3,5] There has been some reported evidence in medical literature with regards to the use of multiple hypothesis testing and the inflation of type I error.[3,6,8–12] As a result, it can indirectly lead to spurious conclusions.[1] In light of the risk of inflation of type I error, there is a number of statistical corrections to account for this detrimental effect.[5,13–21]

In recent years, there has been a growing interest in the orthodontic research and publications due to the advancement in the development of new techniques, treatment modalities and procedures. One of the main aims of the orthodontic research is to evaluate the effectiveness of an intervention in order to achieve a sound conclusion on different treatment approaches.[22] With this, lateral cephalometric radiograph is frequently used in orthodontic research to determine the effectiveness of different orthodontic treatment modalities.[23]

Koletsi et al. reported that the number of published systematic reviews in the orthodontic literature was in an increasing trend with a higher number of interventional studies over the last 15 years.[24] A retrospective review by Gibson and Harrison[25] investigated the types of study published in the four main orthodontic journals- the American Journal of Orthodontics and Dentofacial Orthopedics (AJODO), The Angle Orthodontist (AO), the European Journal of Orthodontics (EJO) and the Journal of Orthodontics (JO) between 1999 and 2008 found that

75% of the clinical based studies were mainly examining diagnosis, development and treatment of the human subjects. The findings were similar to the study by Mavropoulos and Kiliaridis[26] looking at the published orthodontic literature in the last two decades, which found that treatment evaluation and diagnosis were the main scope of orthodontic interest.

Looking at the publication trend on the study types in the last 20 years, it can then be postulated that there is a potential number of clinical trials and publications involving the use of lateral cephalometric radiograph as a tool for diagnosis, treatment evaluation and development of orthodontic appliance. The problem however arises especially with the use of multiple lateral cephalometric variables as the outcome measure. This leads to multiplicity of data when multiple hypothesis testing is performed in the cephalometric study.[27] The issue of multiplicity however has been highlighted by a few authors in the orthodontic literature with some recommendations made to reduce the risk of false positive findings.[21,27–32] In addition, multiple hypothesis testing also arises when comparisons are made between two and or more groups of subjects or changes between different time points using numerous lateral cephalometric variables.

To date, there are no previous published studies looking at multiple hypothesis testing with respect to the use of multiple lateral cephalometric variables as the outcome measure. Therefore, the aim of this research is to examine the extent of the multiple hypothesis testing and its correction in orthodontic literature in relation to the use of lateral cephalometric variables as the outcome measure.

Chapter 2: Literature review

2.1 Hypothesis testing

A hypothesis may be defined as a statement about one or more populations.[2] In statistics, a population represents the entire group of individuals in whom we are interested.[1] An understanding of the basic statistical concepts is paramount in interpreting the results of the hypothesis testing.[28,33,34] The purpose of hypothesis testing is to guide the researchers in achieving a valid conclusion related to the population by drawing inferences from a sample that represent the target population.[1,2]

There are two types of hypothesis which are research hypothesis and statistical hypothesis. The research hypothesis deals with the assumption made from years of observation on the part of researchers that initiate the research. It subsequently leads towards forming the statistical hypothesis. On the other hand, the statistical hypothesis is a hypothesis that is evaluated by appropriate statistical methods. There are two statistical hypotheses in hypothesis testing which are important during a clinical study and which should be stated explicitly when designing a research study.[2]

The null hypothesis [H_0] is the hypothesis to be tested in a research study. It is also known as the hypothesis of no difference.[1,2] The null hypothesis [H_0] is either rejected or not rejected during the statistical testing process. If it is rejected, it shows that the data are unlikely to be compatible with the null hypothesis [H_0], however it may support some other hypothesis. In contrast, if the null hypothesis [H_0] is not rejected, one can conclude that the data do not provide sufficient evidence to cause rejection.[2]

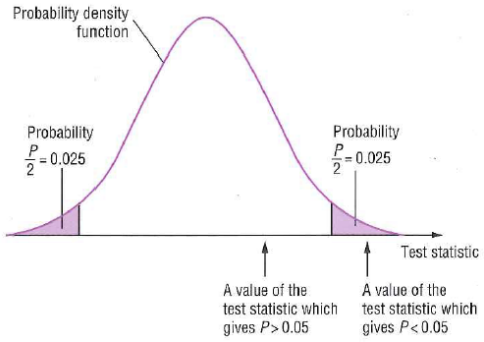
The alternative hypothesis [HA] is a statement which holds if the null hypothesis is not true.[1,2] It relates more directly to the scenario that one wishes to further investigate.[1] Usually the alternative hypothesis [HA] and the research hypothesis are similar and these two terms are used interchangeably.[2]

2.1.1 P value

The definition of the p value is the probability of the observed result, plus more extreme results, if the null hypothesis [H_0] is true.[1] The threshold value for the level of significance, denoted by alpha (α), is arbitrary and usually set in advance.[1–3] It can also be interpreted as a measure of the strength of the evidence from a data set against the null hypothesis [H_0]. Since it is a probability, it has the value between 0 and 1. A p value closer to 0 suggests that the chance of obtaining the observed difference is low, whereas value closes to 1 indicates there

is no difference between the groups. The goal of hypothesis testing using p value is not to 'accept' or 'reject' null hypothesis [H_0]. Rather, it is more of to estimate the likelihood of a real observed difference provided the null hypothesis [H_0] is true.[35]

Figure 2.1 Graphical depiction of the definition of a two sided p value



The curve represents the probability of every observed outcome under the null hypothesis. The p value is the probability of the observed outcome plus all 'more extreme' outcomes, represented by the shaded 'tail area'. Adapted from Petrie and Sabin.[1]

The common cut-off point used for a p value is 0.05.[1–3] Therefore, if the p value is less than 0.05, there is sufficient evidence to reject the null hypothesis [H_0] as there is little chance of the observed results occurring if the null hypothesis [H_0] is true. We then reject the null hypothesis [H_0] and conclude that the results are significant at 5% level.[1] On the other hand, it also suggests that in one out of twenty studies, the null hypothesis is true. Therefore, there is a possibility that the accepted 'significant' result will be wrong 5% of the time.[3,6,36] This is known as a type I error.[1–7]

In contrast, if the p value is equal or greater than 0.05, there is insufficient evidence to reject the null hypothesis [H_0], therefore, the results are not significant at 5% level. It does not conclude that null hypothesis [H_0] is true, however, it is merely a lack of evidence in rejecting the null hypothesis [H_0].[1]

Apart from the significance level of 0.05, the frequently seen values are 0.1, 0.01 and 0.001.[1,2] Under circumstances in which stronger evidence is required before rejecting the null hypothesis [H_0], a p value of 0.01 or 0.001 can be selected.[1,4] The chosen cut-off point for the p value is known as the significance level of the test, which must be decided before the stage of data collection.[1]

2.1.1.1 Problem with p value

It has been recognised that over-dependent on the p values when presenting and interpreting the results in the dichotomy of significant or non-significant is often misleading and unreliable.[35,37–39]

The p value itself is influenced by the sample size and the variances.[40] The p value becomes smaller when there is a larger sample size and a smaller standard deviation. However, smaller p value does not suggest the presence of important clinical effects and in return, larger p value does not advocate a lack of effect.[35,40,41] The p value therefore provides no insight into practical relevance due to the lack of the effect size, range and the clinical importance of the observed results.[38,40,41]

The p value provides limited information on the measure of evidence against a hypothesis. The usual cut-off point of 0.05 is arbitrary and it cannot be considered as absolute. A p value closer to 0.05 indicates a weak evidence to reject the null hypothesis. Equally a larger p value leads to accepting the null hypothesis, however it does not imply that the evidence is in favor of the null hypothesis, it is just purely insufficient evidence for the rejection.[40,41]

It has also been shown that in most of the dental publications, only p values were reported and used to reach the conclusions about the treatment outcome, failing to recognise the importance of the effect size and its range.[42]

2.1.2 Errors in hypothesis testing

2.1.2.1 Type I error

This is known as a false positive error in hypothesis testing. The researcher rejects the null hypothesis [H_0] when it is true and concludes that an effect exists when it does not. It is equivalent to the threshold used for statistical significance, generally 0.05 which is again represented by alpha (α).[1–7]

2.1.2.2 Type II error

This is known as a false negative error in hypothesis testing. The researchers do not reject the null hypothesis [H_0] when it is false and conclude that there is no effect when a true effect exists. The probability of making a type II error is denoted by Beta (β). Its counterpart with the equation $1 - \beta$ is the power of the test. The power is the measure of the possibility of detecting possible difference between groups provided that such a difference exists. Normally, β is arbitrarily set at 0.1 or 0.2, which means a study has either 90% or 80% power to detect a given difference at a specified degree of significance.[1,2,4]

2.2 Multiple hypothesis testing

Multiple hypothesis testing is a common problem in medical literature.[3,6,8,9,18,36,11,43,10,12] In clinical studies, performing multiple statistical tests is frequent when the researchers may wish to compare groups on several different outcomes.[3] It is advantageous as different measured parameters can provide useful data about several aspects of treatment responses. Moreover, secondary outcome analysis can help in the interpretation of the primary outcome measurement.[44]

Nevertheless, when the number of significance tests on a data set increases, there is a greater possibility of finding a false positive result.[1,6,7,9] In other words, the risk of committing a type I error will increase drastically.[1–7] The common situations involving multiple hypothesis testing within a data set are:[1,7,44]

- subgroup analysis to determine differences in treatment outcome in one or more subsets of subjects
- use of multiple predictors in a study
- use of multiple outcomes variables when different endpoints can be used to assess a treatment effect
- multiple treatment comparison for a single outcome variable in three or more treatment groups
- multiple definitions for the exposure and outcomes
- repeated measures over a period of time on the same outcome
- interim analysis of the treatment effect at different stages of treatment during the research trial
- data dredging looking for relationship of different outcome measurement in particular when there is no prior specification on the relationship of specific interest

Lateral cephalometric dataset typically consist of multiple cephalometric variables that measure dento-skeletal and soft tissue changes.[45–47] This set of cephalometric variables are commonly used in a cephalometric study, henceforth it is termed as the lateral cephalometric outcome measurement. A given set of lateral cephalometric dataset is composed of at least 5-10 cephalometric variables; and frequently all are subjected to hypothesis testing for the outcome interest of the study.[21]

Moreover, studies using lateral cephalometric variables also include a set of two group comparisons across multiple outcomes (e.g. differences between the two groups across all cephalometric measures) or multiple group comparisons within an analysis of treatment

changes with time (e.g. differences between several treatment groups and treatment changes with time). Collectively, this is defined as multiple comparisons or multiplicity of data where there is an increased probability of finding a statistically significant result even if the null hypothesis is true, just by chance alone.[19]

2.2.1 Methods for calculation of error rates

Previous studies in the medical field have looked at the error rates and the following statistics are used to quantify the risk of false positive results. Each error rate calculation assumes that the tests are independent and results are presented in probability, number and percentage.[6,8]

2.2.1.1 Family wise error rate

This is the commonly used formula in quantifying the probability of making at least one type I error in a family of hypothesis testing. It is calculated using the formula $1-(1-\alpha)^{\mathrm{C}}$ where α is the significance level and C denotes the number of hypothesis tests. Given an example, if 5 independent tests at 0.05 significance level are performed in which the null hypothesis is true in every case, the probability of at least one test that would be incorrectly rejected is $1-(1-0.05)^5 = 0.23\ (23\%)$. Consequently, the chance of a single false positive for 5 simultaneous tests is 23% which is much greater than the accepted 5%.

2.2.1.2 Error rate per experiment

This is the expected number of type I error in a particular group of statistical significance tests, denoted by $\mathrm{C}(\alpha)$ where C is the number of comparisons and α is the level of significance which is constant across all tests. As an example, given 20 independent hypothesis tests at a significance level of 0.05, this would be $20(0.05) = 1$. It means that one would expect one type I error in 20 statistical tests at 0.05 significance level.

2.2.1.3 Percent error rate

It suggests the percentage of results labelled as statistically significant that are likely to be due to chance with the formula $100\mathrm{C}\propto/\mathrm{M}$, where C is the number of comparisons, α is the significance level for a set of comparisons and M is the number of statistical tests with p value less than the selected significance level. If 3 out of 5 comparisons are statistically significant, this would be $\frac{100(5)(0.05)}{3} = 8.3\%$.

2.2.2 Prevalence of type I error

In reality, it is difficult to estimate the proportion of significant findings that are occurring just by chance alone. It also depends on the study power as the lower the study power, the more likely it is to result in a false positive finding. Therefore, on the basis of varying assumptions, the proportion of the false positive findings has been estimated to range between 1.5% to 96.1% with approximately 50% being the most likely figure.[48]

Nevertheless, there is a small number of published statistical reviews looking at the inflation of type I error associated with multiple hypothesis testing in the medical literature.[3,6,8,9] Ottenbacher[8] looked at five issues of both the American Journal of Public Health and the American Journal of Epidemiology published in 1996. A total number of 173 articles were evaluated. This study reported a high mean family wise error rate in both journals with a score of 0.68 (68%) and 0.70 (70%) respectively, indicating the probability of at least one type I error occurring among these tests. The mean expected number of errors (error rate per experiment) was 0.90 and 0.87 respectively whereas the average percent error rate for the studies in both journals were 19.16% and 18.73% respectively.

An analysis of the use of multiple comparison corrections examining 6415 abstracts was carried out by Stacey et al.[9] with the aim to evaluate the prevalence of multiple testing correction and the percentage of type I error in ophthalmology research. The percentage of family-wise error rate (false positive outcome) was reported in 30% of the abstracts with five or more p value and about half (50%) of the abstracts with ten or more p values.

Walenkamp et al.[3] reviewed the use of multiple hypothesis testing in orthopedic literature looking at two orthopedic journals, the Journal of Bone and Joint Surgery American Edition and the Journal of Bone and Joint Surgery British Edition (Journal A and B) in 2010. The estimated median risk of committing a type I error (family-wise error rate) was 54% in both journals.

More recently, Kirkham and Weaver[6] used a similar method as in the study by Ottenbacher[8] to quantify multiple testing in otolaryngology literature. The result revealed that the mean probability of obtaining at least one false positive in a family of inferences (family-wise error rate) was 0.41 $\pm$ 0.17 (41% $\pm$ 17%). The reported error rate per experiment was 0.61 $\pm$ 0.78. The mean percentage of significance results likely to be false positives (percentage error rate) was 18% $\pm$ 29%.

2.2.3 Multiple hypothesis testing correction

It has been highlighted in the literature on the need to perform statistical corrections if numerous variables are tested using multiple hypothesis tests in order to counteract the potential occurrence of type I error.[5,13,16,18,20,21,49] Surprisingly, the percentage of statistical corrections reported in different medical specialties remains low and it ranges from 6% to 25%.[3,6,11,43,10]

A review by Dar et al.[43] investigating the misuse of statistical tests in the past three decades reported that only a quarter of the studies performed relevant adjustment for multiple hypothesis testing. However, looking at its positive side, the authors highlighted the rise in the statistical correction to compensate for the false positive results from 11.8% in the 60s to 16.4% in the 70s and it further increased to 38.5% in the 80s.

Despite the recommendation of performing statistical correction on multiple hypothesis testing, Kirkham and Weaver[6] reported that only fourteen studies (10%) of the 140 included articles in some way corrected or accounted for multiple hypothesis testing whereas 126 articles (90%) did not account for this problem in the otolaryngology literature. Among the fourteen articles that addressed the issue of multiple hypothesis testing, only eight articles applied a statistical correction. With regards to the types of statistical corrections being employed, five used the Bonferroni method. The other three corrections were the Tukey-Kramer method, the False Discovery Rate method and one which stated a decreased significance level of 0.005 without discussing the use of any methods of statistical correction.

Previous work by Walenkamp et al.[3] examining multiple hypothesis testing problem in the 2010 annals of two orthopedic journals, the Journal of Bone and Joint Surgery American Edition and the Journal of Bone and Joint Surgery British Edition (Journal A and B) concluded that of the 72 studies from Journal A and 55 studies from Journal B included in the review, correction for multiple hypothesis testing in Journal A and B was only reported in eleven articles (15%) and three articles (5.5%) respectively. Bonferroni's method was the preferred method of choice for statistical correction with ten articles from the Journal A and two articles from the Journal B using such a correction method.

The lack of statistical corrections in multiple outcome comparisons was also observed in depression clinical trials where only 5.8% (n=3) of the studies that have accounted for multiple hypothesis testing with all corrections performed using the Bonferroni adjustment.[10] The study focusing on neurology and psychiatry trials also encountered a similar issue with only 25% (n=15) of the studies taking into account the multiplicity issue with six studies using

Bonferroni correction, seven performing other correction methods (Holm, Hochberg-Benjamini, Sidak, Dunnett and sequential adjustments) and two utilising Multivariate Analysis of Variance (MANOVA).[11]

On the other hand, applying correction for multiple hypothesis testing to reduce type I error can result in studies with reduced statistical power which means that there is a reduced probability of rejecting the null hypothesis [H_0] given that null hypothesis is false (type II error). In other words, it reduces the likelihood that the tests will identify the true differences between the groups.[1,2,4]

There is a number of statistical procedures for controlling of type I error. Several correction methods exist such as Bonferroni, Sidak, Benjamini & Hochberg and Holm's for specified multiple hypothesis testing.[5,13–21,50,51] Generally, they can be categorised into two main groups which are single-step and stepwise statistical correction.[6] Bonferroni and Sidak correction[13] are single-step procedures that apply equal correction to all the p values. Both methods are excellent in controlling the type I error rate, however, the pitfall is the reduced statistical power. Therefore, these statistical corrections are not suitable to be used in studies with small sample size.

Thereafter, stepwise procedures are subsequently introduced to control the rate of false positive results whilst at the same time maintaining the statistical power of the study. Stepwise procedure allows sequential evaluation of the hypothesis testing, followed by rejection of the hypothesis based on the outcome of other hypothesis tests.[6]

The Bonferroni-Holm method (sequentially rejective Bonferroni test)[15] is a proposed method of stepwise correction due to its ease of calculation. It was introduced by Holm in 1979. The first step is to perform the tests to obtain the p value. This is then followed by ranking the tests from the smallest p value to the one with the largest p value. The test with the smallest p value will be tested with a Bonferroni correction involving all tests. If the first test is significant, it will then proceed with the second smallest p value and subsequently it is corrected with a correction involving one less test. In the ordered list of the hypothesis testing, once a statistically non-significant result is obtained, all the subsequent hypothesis testing will be declared as non-significant, regardless of how small the p values are. Although this is a well-known method for multiple testing correction among statisticians, however it is not routinely reported in the literature.[5]

Another technique which is used to control type I error is the false discovery rate which was developed by Benjamini and Hochberg in 1995.[17] In this statistical correction procedure, the individual p value is organised in order from the smallest to largest value. The smallest p value has a rank of i=1, followed by the next smallest probability with i=2 and so on. Each of this individual p value is compared to the Benjamini Hochberg critical value with the equation $d \times i/n$, where i is the rank, n is the total number of tests and d is the chosen false discovery rate. In this context, it is a much more conservative procedure as compared to the ordinary method of rejecting hypothesis testing at p value set at 0.05, however it is more powerful than the Bonferroni correction which compares the p value at a similar significance level for all the hypothesis tests.

Among all the correction methods, Bonferroni correction is considered a simple and popular method among clinical researchers. The Bonferroni correction is one of the methods commonly used in the correction of multiple hypothesis testing to avoid the inflation of type I error.[3,6] It is named after the Italian statistician Carlo Bonferroni (1892-1960). This method is popularised by Dunn who described the procedure in his articles in the 60s.[52] It is a popular method and has been widely used in different experimental studies including comparison of multiple groups at baseline, looking at the relationship between different variables and evaluation of more than one end point in the clinical trials.[53]

The Bonferroni adjustment adjusts the p value based on the total number of performed statistical tests with an assumption that the statistical tests are independent. In a simpler way, the alpha level (α) of 0.05 is divided by the number of comparisons being conducted. As an example, if there are 4 statistical tests, the p value threshold would be $0.05 / 4 = 0.0125$ for each of the individual test. The Bonferroni method avoids increasing the likelihood of type I error. However, it becomes overly conservative when the outcome variables are correlated, leading to type II error such that the real differences may not be discovered. In other words, despite the reduction of the number of false positive results, it indirectly raises the number of null hypothesis that are not rejected when in reality they should have been rejected. As a result, it reduces the power of a study in detecting an important effect.[7,27,36,53,54]

Armstrong[36] investigated the use of Bonferroni correction with regards to multiple hypothesis testing in optometric literature and reported that one third of the articles did not make any adjustment on p value for multiple comparisons. Of the 142 articles reviewed, two third of the articles reported a correction on the p value with 9 articles provided a clear rationale for its use e.g. to avoid a type I error, whereas 86 articles provided no discussion on the rationale for the use of the Bonferroni correction. This review indicated that the Bonferroni correction was

used in 51 articles (36%) and the remainder of the correction methods applied being the Bonferroni-Holm method, standard Abbott formula, the false discovery rate, the Hochberg method and other post-hoc procedure such as Scheffe's test.

Ottenbacher[8] recommended the use of Bonferroni adjustment to reduce the chance of making a type I error. However, the author highlighted the loss of statistical power (type II error) when effort is made to reduce the type I error.

There has been much debate in the literature in relation to the use of the Bonferroni method in controlling the type I error.[53] There are several criticisms made of the procedure, particularly by Perneger.[54] In reality, it is considered too conservative resulting in a high level of type II errors. Furthermore, it is a test of a 'universal' null hypothesis [H_0] against an alternative hypothesis [HA]. Given an example, if there are 20 different comparisons between two groups and that all comparisons are significant in all the hypothesis tests. Hence, the 'universal' null hypothesis is rejected. Arguably, this is of little relevance to researcher as the researchers who are more interested in assessing the statistical significance of the individual tests. Thus, there are those who advocated no correction should be made[55,56] and another group who supported the use of correction for multiple hypothesis testing.[5,13,16,18–21,49]

2.2.4 Post-hoc adjustment [Multiple groups within an Analysis of Variance (ANOVA)]

In order to reduce the likelihood of a type I error, an omnibus test such as the F-ratio in ANOVA to reduce the number of comparisons performed can be used. The F-test is known as an omnibus test because it can identify an overall difference among all the groups when comparison is made simultaneously between three or more means using a single test. When all pairwise comparisons are made for n groups, the total number of possible combinations can be calculated using the formula $n \times (n - 1)/2$. Given an example, the total number of pairwise comparisons is six if there are four groups to be compared in a study. The experiment wise error rate which is the probability of at least one type I error is $1 - (1 - 0.05)^6 = 0.26\ (26\%)$, that it is significantly higher than the predetermined level for rejecting the null hypothesis at significance level of 0.05. Thus, the false positive error rate can accelerate beyond the accepted rate of 5% when multiple comparisons are carried out. As a result, ANOVA can be employed to identify the difference in means using a single test rather than to perform multiple separate pairwise comparisons. Hence, an F-ratio less than the significance level would have prevented any further unnecessary testing.[57]

Provided that the F-ratio is significant, indicating that there is a significant difference in the means between groups, a further specific pairwise post-hoc analyses can be performed to reveal the origins of the significance. There are a number of different strategies to control the overall type I error rate for post-hoc analysis which include Tukey's HSD (Honestly Significant Difference), Scheffe's procedure, Bonferroni's procedure, Newman-Keuls procedure and Dunnett's test.[57]

2.2.5 Literature on statistical testing in orthodontic research

In the era of evidence-based dentistry, an understanding of the study design and statistics is essential for proper evaluation of the robustness of the clinical studies. It allows sound clinical decisions to be made for the best interest of the patient.[28,33,34] Therefore, it is important for the orthodontists to grasp a good understanding on the types of statistical analysis used in the orthodontic research particularly to be able to recognise multiplicity issue in a cephalometric study.[27,28]

Statistical tests are performed to assess whether the data set provides a strong evidence to reject the null hypothesis [H_0]. There are two main types of statistical test namely parametric and non-parametric test. Parametric tests require assumptions about data normality which is valid for the populations from which the samples are taken from whereas non-parametric tests are used when the data does not conform to the assumptions made about data normality.[1] The decision on the choice of statistical test relies on the research aims, study design, sample size, the pattern and distribution of the data set and ultimately the outcome measurement of the study.[58]

Summarising the statistics used in the orthodontic literature, Rinchuse et al.[33] observed a dramatic increase in the complexity of study designs and the use of descriptive versus inferential statistics in the American Journal of Orthodontics and Dentofacial Orthopedics (AJODO) articles in the past 25 years. It showed an increasing trend in the use of statistical procedures with the reported percentage of the statistical tests being 43.1%, 75.9% and 94% in 1975, 1985 and 2003 respectively. The observed rise in the percentage of studies using statistics in 1985 (75.9%) was almost double the number seen in 1975 (43.1%), mainly due to an increased use of inferential statistics (23.7% in 1975 to 56.3% in 1985) versus descriptive statistics. In 2003, there were 205 publications with 134 original articles. 4 studies did not use statistics (two essays, a case report and a qualitative analysis) and the remaining 130 articles reported a total of 284 statistical tests. Of the 284 statistical procedures, ten used descriptive statistics, 265 were inferential statistical methods and nine were categorised into miscellaneous group.

Rinchuse et al.[33] and Law et al.[34] both examined the statistical methods of the articles published in AJODO in 2003 and 2008 respectively. They reported that the most commonly used parametric inferential statistical tests in orthodontic literature were Student's T-test, ANOVA and correlation/ regression analysis whereas the most often used non-parametric tests were chi-square test (χ^2) followed by Mann-Whitney U test.

Rinchuse et al.[33] found thirty post-hoc analysis with Bonferroni correction which was the most commonly applied post-hoc test. However, Law et al.[34] observed a relatively different in the use of post-hoc analysis with Tukey adjustment being the most frequently reported procedure. Nonetheless, both studies were not comprehensive and only examined the use of statistics in AJODO articles that were published in the year 2003 and 2008 respectively.

2.3 Cephalometric

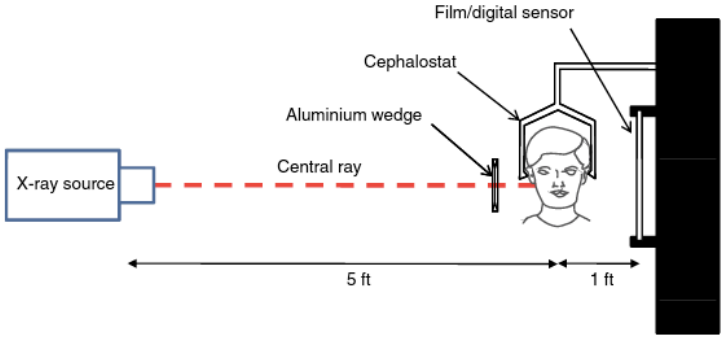
2.3.1 Introduction

Cephalometry is the analysis and interpretation of the relationship of the cranium, facial bones, teeth and surrounding soft tissues using the standardised radiographs.[59] It originally came from the anthropologic work (study of living subjects) and craniometry (dry skulls) where measurements are taken to study the craniofacial forms and structures.[60] Pacini was the first to take a lateral head film as early as 1921 for anthropologic purposes (study of human development, classification and deviations).[61] In 1931, the invention of the cephalostat simultaneously by Broadbent[62] in USA and H. Hofrath[63] in Germany popularised the use of cephalometric radiograph. In clinical orthodontic practice, it is taken in true lateral view.[62,63]

2.3.2 The cephalostat

This radiographic technique focuses on the imaging of the craniofacial region that involves a high powered X-ray machine, a head collar known as a cephalostat (previously known as head-holder or cephalometer), an image receptor system (film cassette) and film cassette holder.[47,59,64,65] Cephalostat helps to stabilise the mid-sagittal plane of the head at a fixed distance from both the X ray source and film using a set of ear posts placed within each external auditory meatus in order to maintain a constant magnification when lateral cephalometric is taken.[47,59,64,65] A locking nasal positioner is placed and secured on the bridge of the patient's nose to support the face and prevent rotation around the ear rods in the sagittal plane.[65] For a lateral cephalometric radiograph, the midsagittal plane of the patient is perpendicular to the x-ray beam and parallel to the film.[47,64,65] The distance between the X ray tube and patient is usually 5 feet and from the patient to the film is 1 foot.[59,64,65]

Figure 2.2 The relationship of the X-ray tube, patient's head and film when taking a lateral cephalometric radiograph



Adapted from Burford and Newell.[64]

2.3.3 Patient positioning

Originally, the patient's head is positioned in the cephalostat which is orientated with the Frankfort plane parallel to the floor as suggested by Broadbent.[62] However, Solow and Tallgren[66] suggested the use of 'natural head position' (NHP) when taking a lateral cephalogram. The NHP represents the true horizontal plane and this position can be naturally obtained when the patient who is standing or sitting in the cephalostat is instructed to gaze at a distance at eye level on the wall/ mirror in front of them.[67] The lateral cephalometric radiograph is taken when the teeth are in the retruded contact position (RCP).[47] To improve the soft tissue outline on the lateral cephalometric radiograph, an aluminium wedge (soft tissue shield) is placed on the film cassette or within the x-ray apparatus to reduce the beam's energy over the soft tissue area.[47,59,65,68]

2.3.4 Magnification

Some degree of magnification is inevitable in any of the radiographs taken.[47,59,65] Therefore, to minimise the magnification error, the distance from the x-ray tube head to the midsagittal plane of the patient is fixed at 5 feet and 1 foot from the patient to the film.[59,65,68] This ensures that the x-ray energy is travelling in a more parallel direction towards the patient/ film, hence the reduction of magnification on the lateral cephalometric radiograph. This fixed distance allows a more consistent measurements to be obtained from the patient that it produces magnification which is consistent but within tolerable limit.[65] The magnification of a lateral cephalometric radiograph ranges from 7% to 8%.[59,64]

A scale is usually seen on a lateral cephalogram to check for magnification and thus allows comparison between different radiographs.[59] Essentially, the effective dose (mSv) for a lateral cephalometric in the United Kingdom (UK) ranges from 0.0022-0.0056.[68]

2.3.5 Uses of cephalometric analysis

The use of lateral cephalogram as a clinical tool has become a standard form of orthodontic care since the development of modern orthodontics by Edward Angle.[45,69] From a historical point of view, the cephalostat was primarily a research tool for studies in the growth and development of the craniofacial complex.[45,60,64] The initial longitudinal study (Bolton study) looking at the development of the craniofacial region using serial cephalometric records begun in 1928 which until today it remains the most extensive source of human growth and development data for clinical and research purposes.[70] Subsequently, Brodie published his PhD thesis in 1940 observing the human growth pattern from birth to the eighth year in which the study was based largely on Broadbent's lateral cephalometric collection.[71] Thus, the documentation of the craniofacial growth from previous research allows orthodontists to understand and appreciate in depth the concept of normal growth and development.[45,69]

Recognising the fundamental role of cephalometrics to the understanding of the craniofacial growth and development, Brodie and the team was the first to publish treatment results based on the cephalometric analysis in 1938.[72] In essence, it then became apparent that lateral cephalometric could be used to assess dento-skeletal proportions and to determine the underlying aetiology of the dental malocclusion for diagnosis and treatment planning.[45,47,59,64] The use of lateral cephalogram enables assessment of the relationship in the craniofacial region in vertical and horizontal dimension. It allows evaluation of the relationships between major functional units of the face which are the jaws to the cranial base, maxilla to mandible, teeth to the supporting bone and the relationship of teeth position to the facial profile.[23,45] As a result, the cephalometric analysis is formed by a combination of different cephalometric variables to yield a description of the relationships of these functional components.[45]

Additionally, superimposition from serial cephalometric radiographs taken before, during and after treatment can be used to study the changes in the craniofacial region and teeth position to evaluate the changes brought about by orthodontic treatment.[45,47] It essentially forms the basis of assessing the effect of orthodontic treatment and is the principal approach for observing treatment response in clinical studies[47]. On the other hand, lateral cephalogram can also be used to evaluate changes obtained from different treatment modalities, thus assessing the effectiveness of the treatment procedures.[23]

The establishment of the normal population norms derived from a number of human population samples using cephalometric analysis has provided useful information on normal average values and standard deviations (SD) for a variety of craniofacial, dento-facial and soft tissue relationship, which are important for orthodontic diagnosis and treatment

planning.[45,47,59,64] Accordingly, there have been a number of different cephalometric analyses being formulated based on different sample sizes and reference groups.[64] For a diagnostic purpose, the norms from the cephalometric analysis provides a means of comparison of the individual's measurement to reveal the differences between the individual's dento-skeletal relationship and those expected dento-facial relationship.[45] Therefore, a detailed analysis allows appropriate treatment planning in determining the most feasible treatment approach.[47]

The possibilities of detecting pathological changes from a lateral cephalometric radiograph should not be underestimated.[45] Occasionally, this lateral view is useful to localise an unerupted impacted tooth,[45,47,59,64] presence of any pathological changes in the craniofacial region such as anomalies or degenerative changes in the cervical vertebrae region[45,64] and to evaluate the size and morphology of the airway.[45,47,64] The cervical vertebrae as seen from the lateral cephalometric radiograph can be beneficial to assess whether there is remaining growth potential in order to consider the most suitable treatment modalities for the patient at the treatment planning stage with a view to achieve an optimal treatment outcome.[45,73]

All in all, the use of lateral cephalometric radiograph has subsequently become one of the most important tools in orthodontics for clinical assessment and diagnosis, treatment planning, as a baseline for monitoring treatment progress, detection of impacted teeth and pathological changes, research and audit purposes.[45,47,59,64]

2.3.6 Types of cephalometric analysis

Since the advent of the lateral cephalometric radiographs, there is a number of cephalometric analytical methods that have been developed as a diagnostic tool to assess hard tissue and soft tissue of the patients' facial structures.[45] The earliest cephalometric analysis is Down's analysis[74], followed by Steiner's analysis[75], Ricketts Analysis[76], Wits appraisal[77], Eastman Analysis[78], McNamara Analysis[79] and many other cephalometric analyses. The main cephalometric analyses are summarised in the table below (Table 2.1).

2.3.6.1 Downs analysis

The first cephalometric diagnosis with clinical application was Downs analysis which was published in 1948.[74] The analysis aimed to describe the basis of the skeletal pattern in the presence of normal occlusion that the assessment was subdivided into skeletal and dental components. His rationale was that if the normal pattern and its range of variation could be described, then the abnormal one could be judged by comparison. Further work by Downs[80] presented an excellent way of plotting the dento-facial pattern on a polygon graph[81] to compare individual's dento-facial type to the mean and variation of the average values. By plotting a

set of value on the graph, it allows a quick quantitative and qualitative assessment of the facial type that an individual conforms to.[80] The analysis developed by Downs has useful clinical implication in today's orthodontic practice, however, these analyses have been replaced by more recent standards.[45]

2.3.6.2 Anteroposterior dysplasia

Likewise, the structures' proportions as determined from the lateral cephalometric analysis can be used to determine the possible aetiology of a malocclusion.[45] Wylie presented a cephalometric analysis of evaluating the antero-posterior relationship of the skeletal pattern based on dividing dimensions along the Frankfort plane into contributing linear components. The term "dysplasia" indicates the random combination of craniofacial parts that might be neither abnormally large nor small, but, when taken together, produce an undesirable combination of parts.[82]

2.3.6.3 ANB angle

The ANB angle introduced by Riedel has a great influence in the world of orthodontics. It is one of the most widely accepted diagnostic measurements and it forms part of the cephalometric analysis developed by Riedel. ANB angle relates the maxilla and mandible to the anterior cranial base and it is commonly used to assess the anteroposterior jaw relationship.[83] Riedel analysis is considered the second major analysis after Down's analysis.[23]

2.3.6.4 Steiner analysis

Steiner analysis[75] was first published in 1953 by Cecil Steiner and many elements of the analysis are still widely used in today's orthodontic practice.[45,47] It offers guides in treatment planning by considering the compromises in the incisor positions in order to achieve a normal occlusal relationship when the ANB angle is not ideal.[84] It also incorporates measurements of arch length and other clinical considerations such as facial profile which enables novice orthodontist to determine whether extractions are necessary during treatment planning stage.[23]

2.3.6.5 Tweed triangle

Tweed analysis[85] was the brainchild of Margolis' research and it only consisted of three measurements. Tweed stated that the mandibular incisors are upright over basal bone which is at approximately 90° angle to the mandibular plane in a normal occlusion. From this postulation, he developed a triangle which was formed by the lower central incisor (LI), mandibular plane (MP) and Frankfort horizontal plane (FH).[85] The 'ideal' Tweed triangle is FH/MP= 25°, LI/MP= 90° and FH/LI= 65°. Therefore, one can estimate whether or not dental extraction of premolars is needed in a particular case.

2.3.6.6 Sassouni analysis

Sassouni analysis[86] was the first analysis focusing on both vertical and horizontal relationships and also the interaction between horizontal and vertical facial proportions. It highlighted the correlation of the horizontal anatomic plane namely the mandibular plane, the occlusal plane, the palatal plane, the Frankfort plane and the inclination of the anterior cranial base, therefore indicating the vertical proportionality of the face. In a well-proportioned face, all the five horizontal planes will converge towards a single point (Point O). Even though this analysis is no longer widely used, the emphasis on the assessment of the vertical facial proportions has an influential impact in today's overall cephalometric appraisal.[45]

2.3.6.7 Bjork analysis/ Jarabak analysis

Bjork developed the first ever analysis taking into account the influence of the cranial base on the facial complex structures.[87] Bjork applied a number of planes to form the 'Bjork polygon' by connecting the points namely Nasion (Na)-Sella (S)-Articulare (Ar)-Gonion (Go)-Gnathion (Gn).[88] The principle of this approach is the relationship of the polygon using three angles which are the saddle angle (N-S-Ar), articulare angle (S-Ar-Go) and gonial angle (Ar-Go-Gn) and the lengths of the sides of the polygon in order to determine the growth.[89] This polygon therefore forms an integral part in superimposition for comparisons and research purposes.[90]

Jarabak analysis[89] was based on Bjork's sample together with another 200 orthodontically treated patients. The most important contribution of this analysis is the interpretation of the polygon to estimate the likely nature of facial growth. The sum of the angles is 396°±6°. An increased angle indicates a clockwise growth rotation or a vertical growth pattern whereas a reduced angle indicates a counterclockwise growth rotation or a growth in a horizontal direction. Additionally, this analysis also assesses the anterior and posterior facial height relationship that enables a prediction of the growth changes in the lower face by calculating the posterior facial height (PFH) to the anterior facial height (AFH) ratio. The ratio should be 62% with the value below the mean suggests a clockwise (backward) growth rotation while the value above indicates an anti-clockwise (forward) growth rotation pattern.

2.3.6.8 Eastman analysis

Eastman analysis is the most widely used analysis in the United Kingdom (UK).[47,64] It was originally the work of Clifford Ballard on a random sample of 250 children and adults from a range of age groups at the Eastman Dental Hospital.[91] This analysis was further developed by Richard Mills[78], and the main components of this assessment are still in common use within the UK today. This is usually supplemented with additional measurements.[47] This is known as the Eastman standard values.[64]

2.3.6.9 Harvold analysis

Harvold analysis is used to assess the severity of the jaw discrepancy. The average length of the maxilla and mandible is calculated based on the samples from the Burlington growth study. Hence, when analysing any given patient's measurement, the difference between the 'unit length' of the maxilla and mandible will indicate the degree of the jaw disharmony.[92] Nonetheless, it does not take into account the vertical position of the jaw in the analysis, in which if the vertical distance is increased, it places the mandible more posteriorly. Also, the position of the teeth has no influence on the Harvold values.[45]

2.3.6.10 Wits analysis

Wits analysis was developed primarily to overcome the limitations of ANB angle in determining the jaw discrepancy. It relates the Point A and Point B in a linear dimension from the occlusal plane to determine the skeletal discrepancy between maxilla and mandible.[77] Similar to Sassouni analysis, it takes into account the vertical and horizontal relationship of the jaws, however the limitation is that it is influenced largely by the dentition that it may not reflect the true underlying jaw disharmony. With this approach, if there is a skeletal discrepancy, this analysis does not distinguish which jaw is at fault.[45]

2.3.6.11 Ricketts analysis

Similar to the previous analysis, Ricketts analysis attempted to determine the relationship of the jaws for aesthetic and function, with the exception that it takes into consideration the effect of facial growth and soft tissue change during treatment planning stage[76]; hence the first cephalometric diagnostic system to project treatment plus growth in treatment planning which is known as the visualised treatment objective (VTO). It was the first cephalometric analysis that allowed clinicians to compare their patients with norms based on age, sex and race.[23] The highlight of the analysis is the inclusion of an aesthetic plane to measure the soft tissue lip position in relation to the nose and chin which is known as the 'Ricketts E plane' which forms the soft tissue analysis.[76] Yet, the main limitation of this analysis is the values of the standard data which are mostly from non-specific samples.[23]

2.3.6.12 Pancherz analysis

Pancherz analysis is a type of grid-based analysis used to determine quantitatively the amount of skeletal and dental changes that have occurred as a result of orthodontic treatment. It was introduced by Hans Pancherz to assess the change within and between the maxilla and mandible using a reference line constructed perpendicular to the occlusal plane superimposed on the SN (Sella-Nasion) line with sella (S) as the registering point.[93]

2.3.6.13 McNamara analysis

McNamara analysis combines the core components of previous analysis (Ricketts and Harvold) with an aim to assess the precise jaw and tooth positions as well as relating the maxilla and mandible in sagittal position to the vertical.[79] The two major advantages are that (1) the use of nasion perpendicular to determine the antero-posterior position of maxilla and mandible which approximates the true vertical line and (2) the norms are based on a well-defined Bolton sample.[45]

2.3.6.14 Holdaway analysis

Holdaway analysis is an analysis focusing on quantification of the soft tissue relationships in order to guide clinicians during treatment planning stage with an aim to achieve a harmonious facial profile after orthodontic treatment.[94] The harmony line (H line) is drawn tangent to the soft tissue chin and upper lip that in a well-proportioned face, it should bisect the nose.[59] H angle as described by Holdaway, is an angle formed between H line to the soft tissue Nasion-Pog line. The ideal measurement is 10° when facial convexity value is 0mm. In essence, as there is an increase in the skeletal convexity, H angle will also follow in an incremental trend if balance and harmony of the face is to be achieved.[94]

2.3.6.15 Bass aesthetic analysis

Bass aesthetic analysis aimed to appraise the facial profile and the optimum position of the dentition within the face by drawing a vertical perpendicular from the aesthetic horizontal line (essentially the true horizontal plane), halfway between the subnasale and point A. This line provides the posterior limit of a harmonious chin in which if it is behind this line, the chin appears to be retrusive. A second vertical line is drawn through the subnasale, thus giving rise to the anterior limit of the chin position. The interpretation of this line is that if it is anterior to the line, the chin will appear protrusive. Thus, this aesthetic analysis allows facial balance to be assessed on an individual basis with main consideration placed on the soft tissue balance. Moreover, the facial profile changes can also be determined when monitoring the effect of the treatment.[95]

Table 2.1 Summary of the main cephalometric analysis

Author	Year
Downs	1948
Wylie	1947, 1952
Riedel	1952
Steiner	1953
Tweed	1954
Bjork	1947, 1954
Sassouni	1955
Ricketts	1960
Eastman	1970
Jarabak	1972
Harvold	1974
Wits	1975
Pancherz	1982
Holdaway (soft tissue)	1983
McNamara	1984
Bass (aesthetic)	1991

2.3.7 Current guideline

Currently, the use of lateral cephalometric radiograph is based on the guideline published by British Orthodontic Society (United Kingdom) in 2015.[68] The indications for its use are as follows:

- Patients with a skeletal discrepancy when treatment is carried out using functional appliances or fixed appliances for labio-lingual movement of the incisors.
- It can be helpful in locating and evaluating any unerupted, malformed or ectopic teeth and to estimate the root length of upper incisor.

Additionally, at present, there is no evidence to support the use of cephalometric radiograph for growth prediction and hence it is not recommended that images are taken for this purpose.[96]

2.3.8 Cephalometric landmarks and variables

Knowledge of the craniofacial anatomy is essential to interpret lateral cephalometric radiograph.[97] Cephalometric landmarks are a series of point located from the oro-facial structures or a constructed point from the intersection of two planes.[45] There are a number of frequently used hard tissue and soft tissue cephalometric landmarks when assessing the lateral cephalometric radiograph.[97]

The commonly used hard tissue landmarks are Sella (S), Nasion (Na), Orbitale (Or), Basion (Ba), Bolton Point (Bo), Anterior nasal spine (ANS), Posterior nasal spine (PNS), Pterygomaxillary fissure (Ptm), Point A (A/ subspinale), Point B (B/ supramentale), Gonion (Go), Gnathion (Gn), Pogonion (Pog), Menton (Me), Porion (Po) and Articulare (Ar). [97]

Whereas the frequently used soft tissue landmarks are glabella (G), Inferior labial sulcus (Ils), Labrale inferius (Li), Labrale superius (Ls), Menton soft tissue (Ms), Nasion soft tissue (Ns), Pronasale (Pn), Pogonion soft tissue (Pos), Superior labial sulcus (Sls), Subnasale (Sn), Stomion (St), Stomion inferius (Sti) and Stomion superius (Sts).[97]

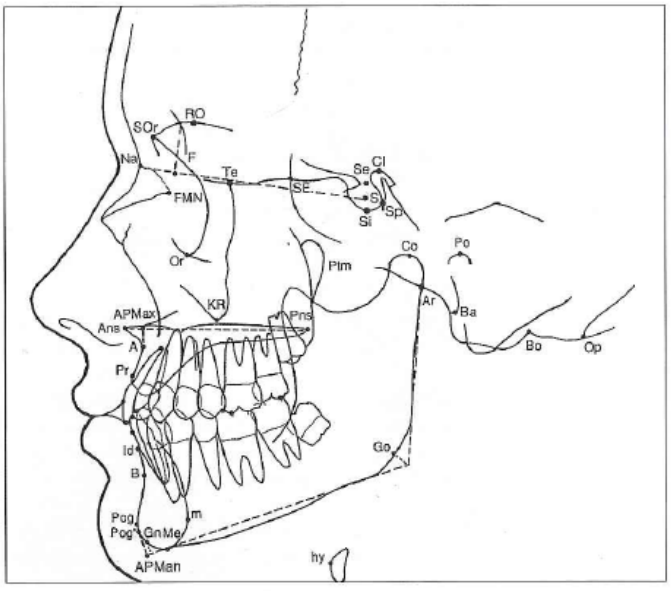
The definition of the hard tissue cephalometric landmarks is as described in Table 2.2.[97]

Table 2.2 Definition of the hard tissue landmarks

No	Landmarks	Definition
1.	Sella (S)	The constructed point representing the midpoint of pituitary fossa (sella turcica).
2.	Nasion (Na)	The most anterior point of the frontonasal suture.
3.	Orbitale (Or)	The lowest point in the inferior margin of the orbit, midpoint between bilateral structures.
4.	Basion (Ba)	The median point of the anterior margin of the foramen magnum, located by following the image of the slope of the inferior border of the basilar part of the occipital bone to its posterior limit.
5.	Bolton point (Bo)	Point in space (roughly at the centre of foramen magnum) that is located on the lateral cephalometric radiograph by the highest point in the profile image of the post-condylar notches of the occipital bone.
6.	Anterior nasal spine (ANS)	The tip of the body anterior nasal spine.
7.	Posterior nasal spine (PNS)	The intersection of a continuation of the anterior wall of the pterygopalatine fossa and the floor of the nose.
8.	Pterygomaxillary fissure (Ptm)	A bilateral teardrop-shaped area of radiolucency, whose anterior shadow represents the posterior surfaces of the tuberosities of the maxilla.
9.	Point A (A/ subspinale)	The point at the deepest midline concavity on the maxilla between the anterior nasal spine and prosthion.
10.	Point B	The point at the deepest midline concavity on the mandibular

	(B/ supramentale)	symphysis between infradentale and pogonion.
11.	Gonion (Go)	The constructed point of intersection of the ramus plane and the mandibular plane.
12.	Gnathion (Gn)	The most anteroinferior point on the symphysis of the chin.
13.	Pogonion (Pog)	The most anterior point of the body chin.
14.	Menton (Me)	The most inferior midline point on the mandibular symphysis.
15.	Porion (Po)	The superior point of the external auditory meatus (superior margin of temporomandibular fossa which lies at the same level may be substituted in the construction of Frankfort horizontal).
16.	Articulare (Ar)	The point of intersection of the images of the posterior border of the condylar process of the mandible and the inferior border of the basilar part of the occipital bone.
17.	Condylion (Co)	The most superior point on the head of the condylar head.

Figure 2. 3 Cephalometric landmarks of the craniofacial skeleton



Adapted from Viteporn and Athanasiou[97]

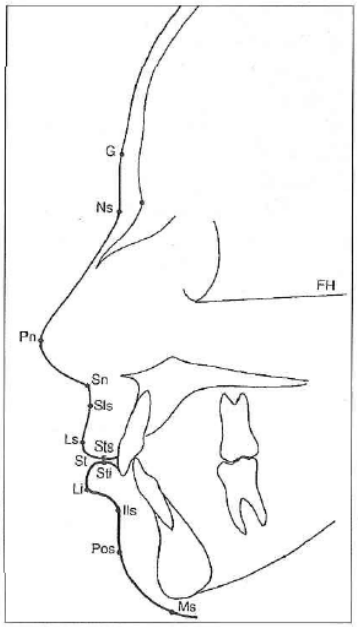
The definition of the soft tissue cephalometric landmarks is as described in Table 2.3.[97]

Table 2.3 Definition of the soft tissue landmarks

No	Landmarks	Definition
1.	Glabella (G)	The most prominent point in the midsagittal plane of forehead.
2.	Inferior labial sulcus (Ils)	The point of greatest concavity in the midline of the lower lip between labrale inferius and menton.
3.	Labrale inferius (Li)	The median point in the lower margin of the lower membranous lip.
4.	Labrale superius (Ls)	The median point in the upper margin of the upper membranous lip.
5.	Menton soft tissue (Ms)	The constructed point of intersection of a vertical co-ordinate from menton and the inferior soft tissue contour of the chin.
6.	Nasion soft tissue (Ns)	The point of deepest concavity of the soft tissue contour of the root of the nose.
7.	Pronasale (Pn)	The most prominent point of the nose.
8.	Pogonion soft tissue (Pos)	The most prominent point on the soft tissue contour of the chin.
9.	Superior labial sulcus (Sls)	The point of greatest concavity in the midline of the upper lip between subnasale and labrale superius.
10.	Subnasale (Sn)	The point where the lower border of the nose meets the outer contour of the upper lip.

11.	Stomion (St)	The midpoint between stomion superius and stomion inferius.
12.	Stomion inferius (Sti)	The highest point of the lower lip.
13.	Stomion superius (Sts)	The lowest point of the upper lip.

Figure 2. 4 Cephalometric landmarks related to the soft tissue profile



Adapted from Viteporn and Athanasiou[97]

Anatomic landmarks on lateral cephalogram are selected and connected to obtain reference plane. The intersection of the two reference planes will give rise to the cephalometric variables. As a result, the cephalometric variables are illustrated as either linear and angular measurements which are expressed in degree (°) or millimeters (mm) to define the skeletal and dental relationships.[46] These cephalometric variables are then used to compare the dento-facial and soft tissue changes between groups of patient.[45–47]

2.3.9 Multiplicity problem with the use of lateral cephalometric variables as the outcome measure

In the orthodontic literature, the multiplicity issue using multiple cephalometric variables as the outcome measures in assessing the dento-skeletal relationships was first mentioned by Tulloch and the team in the United States. They conducted a cephalometric study using a randomised clinical trial (RCT) study design to examine the advantage of early treatment for class II malocclusion. Hence, the authors clearly stated that a limited set of cephalometric measurements were employed in order to reduce the risk of false positive results.[31,32]

The multiplicity problem was further recognised by O' Brien and the team in the United Kingdom (UK) who carried out a similar multi-center randomised controlled trial (RCT) using cephalometric variables as the outcome measures. The method of counteracting this multiplicity problem was comparable to previous studies where the authors restricted assessment to a set of important cephalometric variables when assessing skeletal and dental changes so that it lowered the likelihood of finding a positive result by chance alone.[29,30]

Harrison further highlighted this multiplicity problem in orthodontics especially with the use of multiple cephalometric measurements for comparisons. The following recommendations were made to reduce the chance of a false positive result when conducting orthodontic research that deals with multiple cephalometric measures that are tested for significance. The author suggested to limit the number of cephalometric variables used in a study, to specify the primary outcome measure at the protocol stage of the study and to present the original p value, rather than just presenting it as either $p > 0.05$ or $p < 0.05$.[28]

Pandis observed similar problem as mentioned by previous authors that the interpretation is often based on the p value which might provide the wrong impressions of the treatment outcomes and effectiveness. Therefore, the author recommended either an overall combined end point limited to the area of interest (maxilla, mandible and dentition) or p value correction using Bonferroni method to reduce the chance of false-positive findings.[27]

Simas et al. commended statistical correction to control for false positive rates in dental research. In this article, the authors gave two examples of statistical correction procedures when multiple comparison is made on a set of cephalometric variables. The authors described two statistical correction methods which were Bonferroni correction and Benjamini and Hochberg formula (to control the false discovery rate). The key message from this publication is that multiple outcome comparisons require a carefully selected and correct statistical analysis with appropriate statistical correction in order to decrease a false positive error.[21]

Similarly, Martins and Buschang recommended the use of resampling methods or Bonferroni correction if it is associated with a large number of comparisons.[98]

In view of the multiple hypothesis testing within the cephalometric data set, therefore, a review of the multiplicity problem using cephalometric variables as the outcome measures would appear to be justifiable.

Chapter 3: Study aim and objectives

3.1 Study aim

The aim of the study was to:

- Examine the extent of the multiple hypothesis testing and its correction in orthodontic research in relation to the use of lateral cephalometric variables as the outcome measure.

3.2 Study objectives

The objectives of the study were to:

- Quantify studies with multiple testing in orthodontic research specifically with the use of lateral cephalometric variables as the outcome measure.
- Determine the potential prevalence of false positive results (type I error) in the sample of published articles in orthodontic research related to the use of lateral cephalometric variables as the outcome measure.
- Determine the frequency by which multiple testing are correctly addressed in the statistical analysis.
- Describe methods used for correction of multiple hypothesis testing.
- Determine the association of study type, journal classification, region of authorship, number of researchers and statistician authorship to the application of multiple hypothesis testing correction.
- Compare the electronic search methods with handsearching as the gold standard.

Chapter 4: Methodological Framework

4.1 Study design

This was a retrospective, observational study looking at a sample of published orthodontic articles over a two-year period from 1st January 2014 to 31st December 2015.

4.2 Study selection criteria

To be included in the study, the articles were to meet the following criteria using a PICO format which requires consideration of 4 key components when formulating questions and search strategies.[99] The acronym PICO stands for:

P- Patient, population or problem: characteristics of the patient or population and condition or disease of interest

I- Intervention: intervention used for the patient or population

C- Comparison: alternative to the intervention, if relevant

O- Outcome: the outcome of interest of the study

These were the criteria when considering studies for this review:

- Language: restricted to English language only when searching the electronic databases
- Types of studies: observational or interventional research
- Types of participants (P): patients undergoing orthodontic treatment
- Types of interventions (I): any types of orthodontic treatment modalities; either fixed appliance, removable appliance, combined orthodontic-orthognathic treatment approach or a combination of any orthodontic appliance therapy
- Types of comparisons (C): at least two comparison groups or measurement time points
- Types of outcome measures (O): lateral cephalometric variables in dento-facial region, viewed on a 2-dimensional lateral cephalometry

The following articles were excluded:

- Animal and laboratory studies
- Studies using 3D cephalometry taken from cone beam computed tomography (CBCT)
- Unpublished studies
- Case reports and case series
- Systematic review and meta-analysis
- Letters to editors, book chapters, abstracts or commentaries

- Duplicative studies originating from the same subjects by the same investigators but published in different journals
- Studies with insufficient information or unclear statistical analysis method

Based on previous studies, multiple hypothesis testing was defined as having five or more p values obtained from comparing two or more groups or two or more time points on a set of variables, using separate statistical tests.[3,6]

4.3 Search methods for identification of studies

4.3.1 Electronic searching

For identification of studies, four major electronic databases namely PubMed, Ovid Medline, Scopus and EBSCO Dentistry & Oral Sciences Source were electronically searched using the following search sequence of medical subject headings (MeSH) terms: (orthodontic* OR "orthodontic treatment" OR "orthodontic appliance*") AND (cephalometr* OR "lateral cephalometric") AND (compar* OR analys* OR measure* OR calculat*) to identify relevant articles published from 2014 to 2015.

4.3.2 Handsearching

Additionally, four main orthodontic journals which were the American Journal of Orthodontics and Dentofacial Orthopedics (AJODO), The Angle Orthodontist (AO), the European Journal of Orthodontics (EJO) and the Journal of Orthodontics (JO) were hand-searched systematically for all articles published between 2014 to 2015. This was to identify any relevant articles which were missed from electronic searching. Also, the list of the retrieved publications was cross-checked to avoid inadvertent omission. The references of the included publications were checked for identification of further studies.

4.4 Pilot study

Prior to the commencement of the article search, SCP discussed with the research supervisors (GB, NF) the article selection following title and abstract screening and also data to be extracted from the included studies. A pilot study was undertaken using a specifically designed title and abstract screening form and data extraction form on several papers with the supervisors (GB, NF) in order to improve the content of the forms. The title and abstract screening form and data extraction form were finalised through discussion with supervisors (GB, NF) and were subsequently used in the present study (Appendix 1 and 2).

4.5 Selection process

The initial electronic search was carried out by SCP independently using the MeSH terms. The search results from the databases were compiled using a bibliographic software, Mendeley Desktop (Version 1.17.10, Year 2008-2016, Mendeley Ltd., London, United Kingdom) and any duplicate studies were merged.

After initial piloting and reliability testing, SCP independently performed screening of the titles and/ or abstracts against the predetermined inclusion and exclusion criteria using a specially designed title and abstract screening form (Appendix 1) to identify potentially relevant papers. However, if a decision could not be made based on the title, the abstracts were examined. A maximum of 10 papers were assessed at any one time with a view to prevent errors due to fatigue.

The title and abstract screening form (Appendix 1) contained the following items:

- Title of the paper
- Use of 2-dimensional or 3-dimensional lateral cephalometry
- Evidence of the use of lateral cephalometric variables either in dento-facial or non dento-facial region
- Number of groups/ measurement time points for comparison

The full text research papers were retrieved and assessed accordingly either electronically or in paper format for study eligibility in which multiple hypothesis testing existed. The uncertainties on the study eligibility was discussed with supervisors (GB, NF) until a consensus was reached. All the full text files were added to the entries in the Mendeley Desktop software for data extraction.

4.6 Data extraction and items

A structured data extraction form (Appendix 2) was used to systematically collect the information needed. Each article was assessed on the following items:

- Title of the paper
- Year of publication: 2014 or 2015
- Region of authorship: according to the continent of location of the first author and three categories were created (Americas, Asia and other or Europe)
- Source of journal publication: classification was made based on 2015 SCImago Journal and Country Rank (SJR) indicator[100] and two categories were formed (main orthodontic journal or non-main orthodontic journal)

- Type of study design: retrospective or prospective
- Number of groups (inter or intra-group)/ measurement time points for comparison
- Number of subjects
- Number and list of lateral cephalometric variables
- Level of significance: p value and adjusted p value (if any)
- Each significant p value
- Methods of multiple hypothesis testing correction and its rationale for correction
- Involvement of a statistician/ epidemiologist: association with statistician/ epidemiologist was determined by the affiliation information for the authors (yes or no)
- Primary and/ or secondary outcome measures (if any): yes or no
- Conclusions made according to the stated primary and/ or secondary outcome measures: yes or no

Data extraction was performed independently by SCP using the specifically designed data extraction form (Appendix 2). Accordingly, a maximum of 2 papers were examined at any one time in order to prevent mistakes due to fatigue.

The uncertainties in determining the exact number of statistical tests performed and the number of tests found to be statistically significant from some of the included articles due to the use of complicated statistical analysis and data presentation were discussed with supervisors (GB, NF). Additional input was sought with the value agreed by at least two was used in the data analysis.

4.7 The correction experiment

The correction experiment aimed to provide an overview of the likely prevalence of the false positive results if multiple testing had not been accounted for in the study sample. The use of lateral cephalometric variables as the outcome measure was considered as primary analyses if the families of tests was stated in the aims and objectives of the study, and those that did not were considered as secondary analyses. Therefore, only those studies using lateral cephalometric variables as the primary analyses were included in this correction experiment. For each primary families of statistical testing, the total number of comparisons and all significant p values were recorded. Bonferroni correction was then applied to the significant p values within each of the family of interferences to yield the corrected results. However, in cases where the exact p value was not given and it was displayed as less than a cut off value (e.g. $p < 0.01$), the p value below the cut off figure was recorded at the next significant digit which was P= 0.009.

4.8 Reliability

4.8.1 Title and abstract screening

The level of agreement on the eligibility of the articles between examiners (SCP, GB) was assessed prior to the main study. This was to ensure that an acceptable level of intra-examiner reliability was maintained over time when title and abstract screening was carried out independently by SCP.

For inter-rater reliability assessment, a list of five percent of the journal samples from both electronic and hand searching were prepared using a random number generator[101] by SCP. SCP and GB completed the title and abstract screening independently and in duplicate according to the inclusion and exclusion criteria to every article. If the level of agreement was low between the examiners (SCP, GB), further discussion was anticipated, followed by reassessment after one month on the same papers until a good level of agreement was obtained.

For intra-rater agreement, assessment was carried out after completion of the screening of the title and abstract for the entire sample. Similarly, a list of five percent of the journal articles from both electronic and hand searching were prepared using a random number generator[101] by SCP. The inter-rater and intra-rater reliability were tabulated and assessed using Kappa statistic and percentage agreement.

4.8.2 Data extraction

As for the reliability on the data extraction, a ten percent sample of the included articles were prepared and assessed. The same articles were reassessed by SCP after one month into the data collection period to assess intra-examiner reliability. The intra-rater reliability was tabulated and assessed using Kappa statistic and percentage agreement.

4.9 Data entry

The data were entered into two customised Microsoft Excel spreadsheets (Version 15.14, Year 2015, Microsoft, Microsoft Office 2015, Microsoft Corporation, Redmond, USA) with one customised for the title and abstract screening and another for the data extraction (Appendix 1 and 2).

4.10 Quality assessment

During the stage of data collection, there were no attempts made to assess the quality of the individual articles from the study sample. This was considered to be out of the remit of the aim and objectives of the research to make further evaluation of this aspect.

4.11 Statistical methods

Descriptive statistics were used to analyse the characteristics of the articles, including the total number of included articles published in 2014 to 2015, study type, journal classification, region of authorship, number of researchers and methods of statistical correction. Values were presented as mean ± standard deviation (SD), median and interquartile range (IQR) and range for continuous data and percentages for dichotomous data in tabular form. Chi-square test (χ^2) was used to access the association of study type, journal classification, region of authorship, number of researchers and statistician authorship to the application of multiple hypothesis testing correction. Statistical significance was set at 0.05.

Additionally, each individual article was examined to determine the family-wise error rate, error rate per experiment and percent error rate. Each error rate calculation assumed that the tests were independent of one another within a family of statistical tests.

4.12 Statistical analysis

This was undertaken by using IBM SPSS Statistics, Version 24.0 (Armonk, NY: IBM Corp).

4.13 Ethical implication

This was a retrospective observational study using the secondary data from previous published research. Since there was no direct contact made with study subjects and no identifiable data that was used, therefore ethical consideration was considered to be unnecessary.

Chapter 5: Results

The results are presented in seven sections:

5.1 Results of the search

5.2 Characteristics of the included articles

5.3 Summary of the methods of correction for papers that have accounted for multiple hypothesis testing

5.4 Error rates calculation for articles with unaccounted multiple hypothesis testing

5.5 The correction experiment

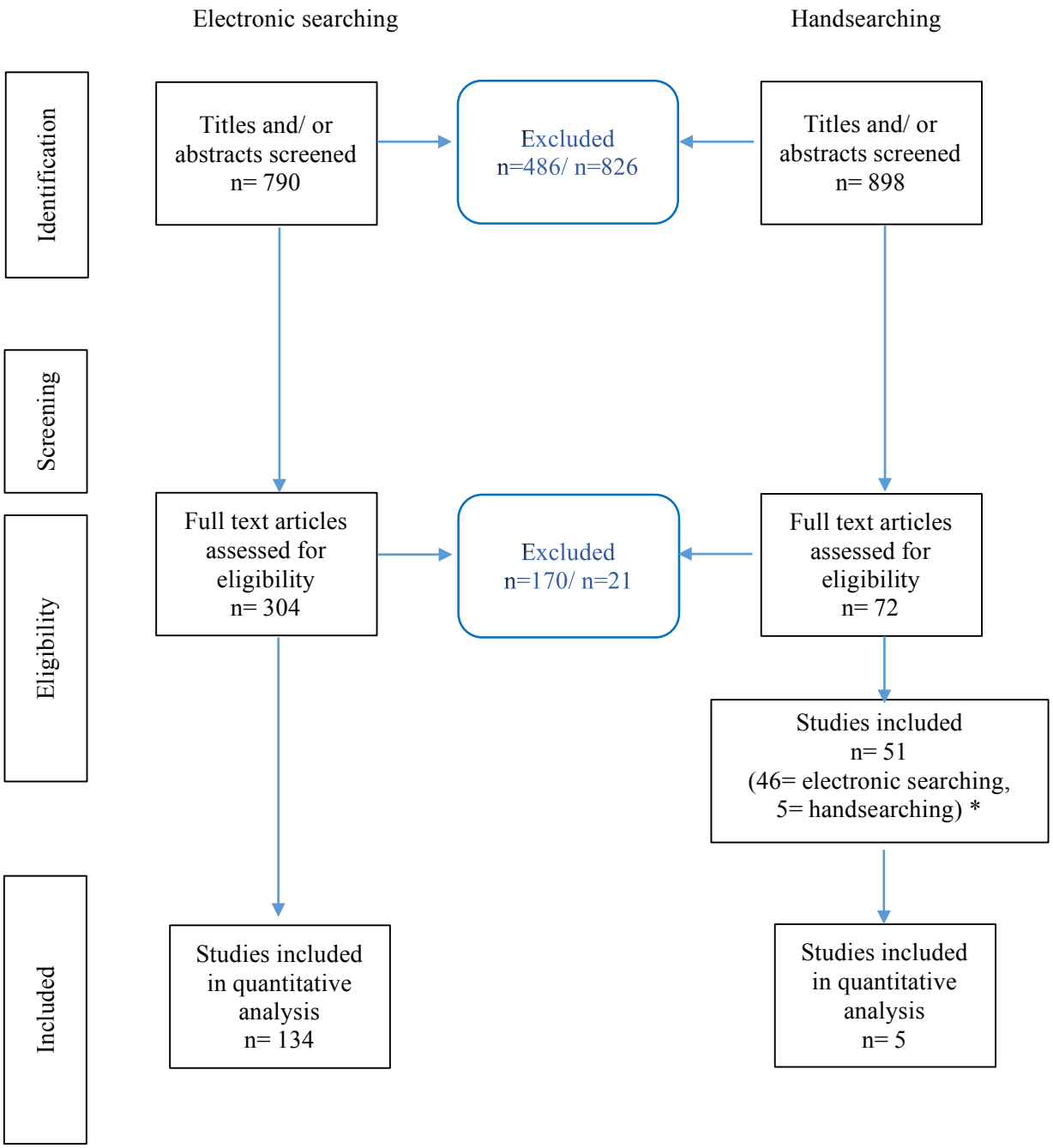
5.6 Factors influencing the multiple hypothesis testing correction

5.7 Inter and intra-reliability testing

5.1 Results of the search

The flowchart indicating the search result is shown in Figure 5.1.

Figure 5.1 Flowchart indicating the search result



* There was an overlapping of papers found from both the electronic and handsearching methods. Of the 51 papers included from the handsearching, 46 papers were also found following the electronic searching.

5.1.1 Overall number of the articles identified

A total of 1688 articles were identified from both electronic and handsearching (790 from electronic searching and 898 from handsearching) between 1st January 2014 and 31st December 2015, as seen in Table 5.1.

Table 5.1 Overall number of the articles identified from both electronic and handsearching

Search method	Number of articles in 2014	Number of articles in 2015	Total number of articles
Electronic searching	441	349	790
Handsearching	452	446	898
Electronic+ Handsearching	893	795	1688

5.1.2 Overall number of the articles fulfilling the eligibility criteria

Following the title and abstract screening against the pre-determined inclusion and exclusion criteria, there was a total of 376 papers that fulfilled the eligibility criteria, as illustrated in Table 5.2. Of these, 304 publications were from electronic searching and 72 were from handsearching. Hence, full text papers were retrieved and assessed appropriately to identify cases of multiple hypothesis testing. Articles were thoroughly reviewed for tables and/ or graphs that were presented with a minimum of five or more p-values.

Table 5.2 Overall number of the articles fulfilling the eligibility criteria

Search method	Number of articles 2014	Number of articles 2015	Total number of articles
Electronic searching	162	142	304
Handsearching	39	33	72
Electronic+ Handsearching	201	175	376

5.1.3 General characteristics of the papers from handsearching

All issues of the AJODO, AO, EJO and JO published in 2014 and 2015 were handsearched. A total of 898 articles were identified in 60 issues of the journals, as demonstrated in Table 5.3, with a total of 72 papers fulfilling the eligibility criteria. Full text papers were then retrieved and comprehensively reviewed in particular looking at the tables and/ or graphs displaying at least five or more p-values. This led to a total of 51 papers that were included in the final analysis of this review. Of the included 51 papers, it was then followed by comparison to the papers found from electronic searching to determine any papers that were missed from electronic searching.

Table 5.3 Overview of the studies characteristics from handsearching

Journal	**Number of issue**	**Number of articles in 2014**	**Number of articles in 2015**	**Total number of articles**	**Number of papers fulfilling the eligibility criteria**	**Included in final analysis**
AJODO	27	173	182	355	16	12
EJO	12	92	94	186	17	12
AO	12	144	139	283	38	26
JO	9	43	31	74	1	1
Total	60	452	446	898	72	51

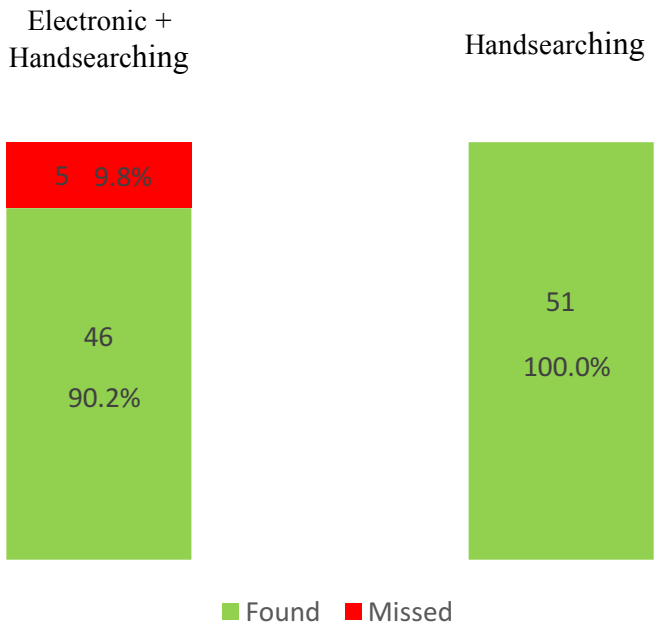
5.1.4 Identification of papers from both the electronic and handsearching

51 papers that were found following handsearching of the four orthodontic journals (AJODO, AO, EJO and JO) were compared to those journal articles found from the electronic searching. Of the included 51 papers from handsearching, 46 publications (90.2%) were also found from the electronic databases. Therefore, handsearching of four orthodontic journals resulted in 5 additional papers (9.8%) which would potentially have been missed if only electronic searching was performed, as shown in Table 5.4, Figure 5.2.

Table 5.4 Number of papers found and missed from electronic and handsearching

Search method	Found (%)	Missed (%)	Total (%)
Electronic searching	46 (90.2%)	5 (9.8%)	51 (100%)
Handsearching	51 (100%)	0 (0%)	51 (100%)

Figure 5.2 Number of articles found and missed by electronic and handsearching



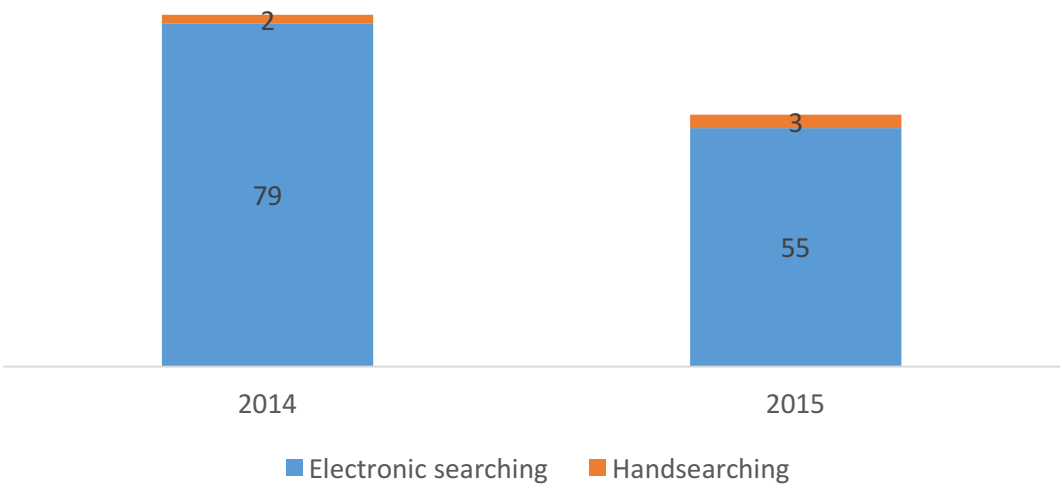
5.1.5 Number of papers associated with multiple hypothesis testing that were included in the final analysis

Of the 376 papers that fulfilled the eligibility criteria, following application of the predefined inclusion and exclusion criteria, a total number of 139 published articles (134 of electronic searching and 5 of handsearching) met the criterion for multiple hypothesis testing, as seen in Table 5.5, Figure 5.3.

Table 5.5 Number of papers included over the two-year period (2014-2015)

Search method	Number of articles in 2014	Number of articles in 2015	Total number of articles
Electronic searching	79	55	134
Handsearching	2	3	5
Electronic + Handsearching	81	58	139

Figure 5.3 Number of included papers published over the two-year period (2014-2015)



5.2 Characteristics of the included articles

The number of subjects, number of distinct families of tests (per article) and number of hypothesis tests (per family of tests) were calculated for each included article. Given an example, if there was an intergroup comparison at three different time points to assess the treatment changes with time, namely T1, T2 and T3 using a list of sixteen cephalometric variables, therefore, three distinct families of sixteen hypothesis tests were recorded.

In cases where there was discrepancy in the number of cephalometric variables (number of hypothesis tests) in each family of tests, the decision was made that the family of tests with the highest number of the cephalometric variables was recorded. This aimed to demonstrate the highest number of hypothesis tests used in a study with a view not to underestimate the error rates to be calculated in this study.

The mean ± SD, median (IQR) and range for the number of subjects, number of families of tests (per article) and number of hypothesis tests (per family of tests) are displayed in Table 5.6. Of the 139 included publications, the number of subjects ranged from 9 to 191 with a mean of 44 participants in a study. The families of tests (per article) ranged from 1 to 9 with a mean of 3 families of tests in a cephalometric study using lateral cephalometric variables as the outcome measure. The number of hypothesis tests (per family of tests) ranged from 5 to 47 with a mean of 20. This indicated that there was an average of 20 cephalometric variables that were employed in a family of tests and with 47 being the highest reported number of lateral cephalometric variables used in a cephalometric study.

Table 5.6 Characteristics of the articles (n= 139)

Characteristics	**Mean± SD**	**Median (IQR)**	**Range**
Subjects	44.32 ± 27.57	39 (28-56)	9-191
Families of tests (per article)	2.83 ± 1.92	3 (1-4)	1-9
Hypothesis tests (per family of tests)	19.63 ± 9.79	18 (12-28)	5-47

5.3 Summary of the methods of correction for papers that have accounted for multiple hypothesis testing and the rationale for its correction

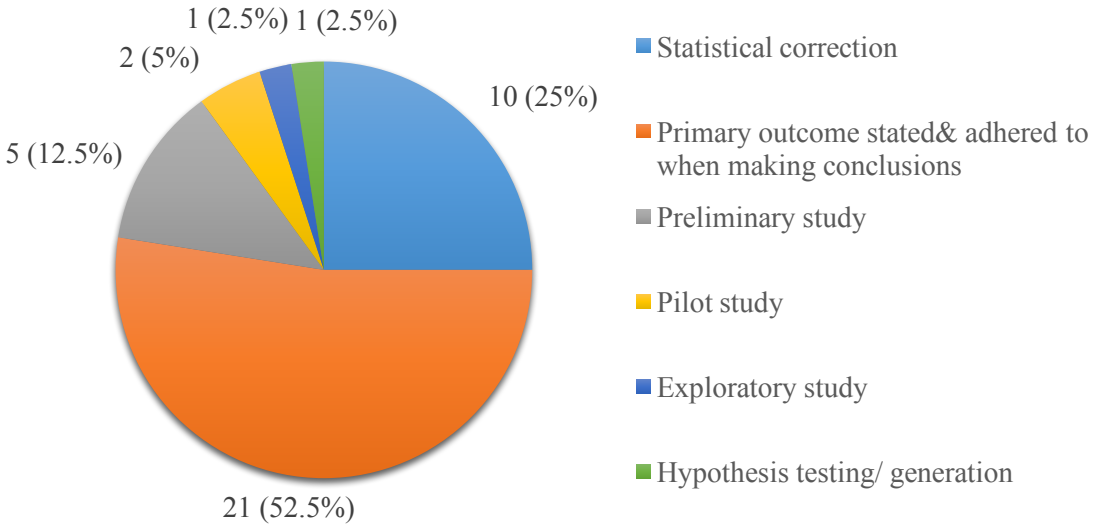
Of the 139 included studies, only 40 papers (29%) in some way corrected or accounted for multiple hypothesis testing, whereas 99 (71%) did not. Of the 40 articles that addressed the problem of multiple hypothesis testing, ten applied statistical correction (25%), twenty-one pre-specified a primary outcome which was adhered to when making conclusions from the study (52.5%), five claimed to be preliminary studies (12.5%), two were pilot studies (5%), one was stated as an exploratory study (2.5%) and one study was aimed at generating hypotheses (2.5%).

Of the ten using statistical correction, seven used Bonferroni correction, two applied Benjamini Hochberg and one chose a decreased significance level of 0.01 without specifying any correction methods, as can be seen in Table 5.7, Figure 5.4.

Table 5.7 Studies in some way corrected or accounted for multiple testing (n=40)

Methods	**Number of papers (%)**
Statistical correction	10 (7 Bonferroni, 2 Benjamini Hochberg & 1 decreased significance level of 0.01) (25%)
Primary outcome stated and adhered to when making conclusions	21 (52.5%)
Preliminary study	5 (12.5%)
Pilot study	2 (5%)
Exploratory study	1 (2.5%)
Hypothesis testing/ generation	1 (2.5%)

Figure 5.4 Distribution of studies that in some way corrected or accounted for multiple testing



The rationale for the statistical correction of multiple testing was examined, as outlined in Table 5.8. Of the ten studies with statistical correction to account for multiple hypothesis testing, four (40%) did not provide a clear rationale or discussion, while six (60%) stated a clear rationale for its use e.g. for multiple testing and to control the type I error.

Table 5.8 Rationale for the statistical correction (n=10)

Rationale	**Number of papers (%)**
For multiple testing	3 (30%)
To control type I error	3 (30%)
No rationale	4 (40%)

5.4 Error rates calculation for articles with unaccounted multiple hypothesis testing

The mean± SD, median (IQR) and range for each error rate calculation are illustrated in Table 5.9. For each error rate calculation, the quantification was expressed in the following order: probability, number and percentage. The mean probability of making at least one type I error (false positive) in a family of inferences (family-wise error rate) was 0.58 ± 0.19 (58% ± 19%). The mean expected number of type I error (false positive) in a particular group of statistical significance tests (error rate per experiment) was 0.97 ± 0.51. The mean percentage of results labelled as statistically significant that were likely to be by chance alone (percent error rate) was 13.44% ± 11.93%.

Table 5.9 Descriptive information for the error rates for articles with unaccounted multiple testing (n=99)

Characteristics	Mean± SD	Median (IQR)	Range
Family-wise error rate	0.58± 0.19	0.60 (0.43-0.76)	0.23-0.91
Error rate per experiment	0.97± 0.51	0.90 (0.55-1.40)	0.25-2.35
Percent error rate (%)	13.44± 11.93	10 (7.63-14.72)	5-100

5.5 The correction experiment

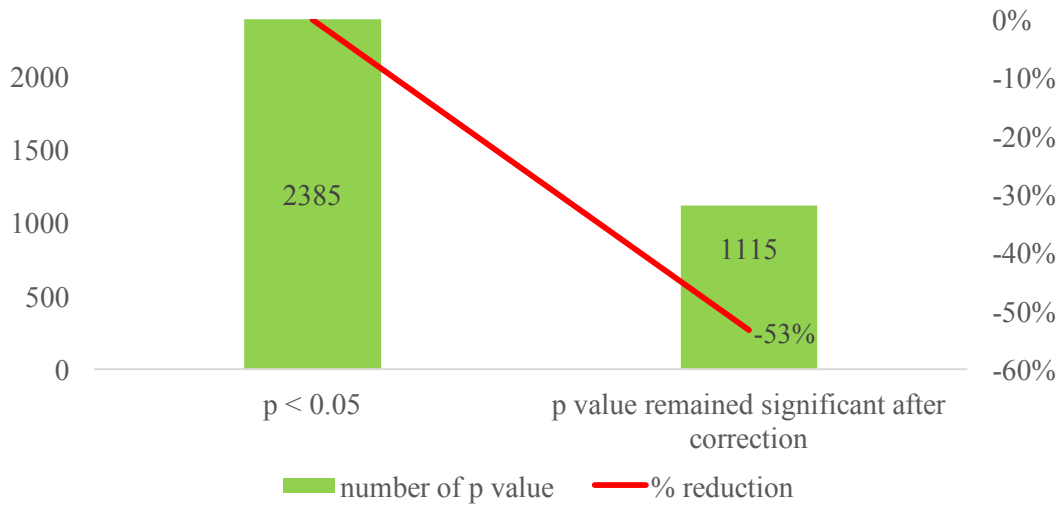
This was only an estimation of the numbers of p values that might potentially become non-significant after the correction experiment because some studies did not provide the exact p value for each hypothesis test. The decision for the cut-off point on the p value without an exact figure was discussed in Section 4.7.

Of the 99 studies with unaccounted multiple testing, 93 articles (94%) were considered to be using lateral cephalometric variables outcome measurements as the primary analyses of the research. The application of the Bonferroni correction affected the results of 86 of the 93 articles (92%) which tested five or more hypotheses, in that it resulted in a change of at least one p value which met statistical significance. Among all the families of tests with multiple hypothesis testing, there was a total number of 2385 significant p values. The Bonferroni correction experiment resulted in only a total of 1115 p values that remained significant, leading to a 53% reduction in the reported significant p values, as shown in Table 5.10, Figure 5.5.

Table 5.10 The correction experiment using the Bonferroni method (n=93)

Characteristic	Total number
p value < 0.05	2385
p value remained significant after Bonferroni correction	1115
Reduction of significant p value (%)	53%

Figure 5. 5 Percentage reduction of the significant p values after the correction experiment with the Bonferroni method



5.6 Factors influencing multiple hypothesis testing correction

The distribution of studies associated with multiple hypothesis testing with and without correction based on study type, journal classification, region of authorship, number of researchers and statistician/ epidemiologist involvement are outlined in Table 5.16.

5.6.1 Study type

Of the 139 included articles, retrospective studies made up 62% (n=86) of all articles published over the two-year period from 2014 to 2015. This in return resulted in a slightly higher proportion of retrospective research (n=22) that had accounted for multiple testing correction when compared to prospective studies (n=18). However, the study type was not associated with the application of a correction for multiple testing (p= 0.289) (see Table 5.11).

Table 5.11 Number of studies based on the study type with and without multiple testing correction

Study type	Number of studies (n=139) (%)	Multiple testing correction (n=40) (%)	No multiple testing correction (n=99) (%)	p value
Prospective	53 (38.1%)	18 (34.0%)	35 (66.0%)	0.289
Retrospective	86 (61.9%)	22 (25.6%)	64 (74.4%)	

5.6.2 Journal classification

The included publications were classified into two main categories which were articles published in the main orthodontic journals and non-main orthodontic journals. The classification of orthodontic journal was based on the 2015 SJR indicator.[100] The top four ranking journals based on the SJR indicator were classified as main orthodontic journals. These included AJODO, AO, EJO and Korean Journal of Orthodontics (KJO) with the following SJR indicator of 1.343, 1.313, 1.129 and 0.982 respectively.

Studies with multiple testing were published more in the non-main orthodontic journals (59%) when compared to the main orthodontic journals (41%). Nonetheless, multiple testing was correctly accounted for in 25 studies published in the main orthodontic journals (43.9%) and only 15 studies in the non-main orthodontic journals (18.3%). The differences were statistically significant (p=0.002), as depicted in Table 5.12.

Table 5.12 Number of studies based on the journal classification with and without multiple testing correction

Journal classification	**Number of studies (n=139) (%)**	**Multiple testing correction (n=40) (%)**	**No multiple testing correction (n=99) (%)**	**p value**
Main orthodontic journal	57 (41%)	25 (43.9%)	32 (56.1%)	0.002
Non-main orthodontic journal	82 (59%)	15 (18.3%)	67 (81.7%)	

5.6.3 Region of authorship

The continent of the authorship was subdivided into the following:

1. Americas
2. Europe
3. Asia/ Others

If the published studies had authors from more than one country, only the country of origin of the first author was recorded. The number and percentage of articles published in each region can be seen in Table 5.13.

Overall, a higher proportion of studies with multiple testing were carried out by the authors in Asia/ others (56%), followed by the authors from Europe (24%) and Americas (20%). In return, almost half of the articles published by Asia/ others authors (n=21) had accounted for multiple hypothesis testing, this was followed by the authors from Europe (n=14) and Americas (n=5). There was no association between the region of authorship to the correction associated with multiple hypothesis testing (p=0.093).

Table 5.13 Number of studies based on the region of authorship with and without multiple testing correction

Region of authorship	**Number of studies (n=139) (%)**	**Multiple testing correction (n=40) (%)**	**No multiple testing correction (n=99) (%)**	**p value**
Americas	28 (20.1%)	5 (17.9%)	23 (82.1%)	0.093
Europe	33 (23.7%)	14 (42.4%)	19 (57.6%)	
Asia/ Others	78 (56.2%)	21 (26.9%)	57 (73.1%)	

5.6.4 Number of researchers

The number of researchers was grouped into the following categories:

1. One to four
2. Five to seven
3. Eight or more

As a whole, more than half of the studies (54%, n= 75) with multiple testing were carried out by one to four authors, followed by five to seven authors (41.7%, n= 58) and eight or more authors (4.3%, n=6).

Of the 40 studies that have accounted for multiple testing, 33.3% (n=25) of the papers involved one to four authors, followed by 24.1% (n=14) of the publications with five to seven researchers and 16.7% (n=1) of the studies with eight or more authors. The number of researchers was not associated with the application of multiple testing correction (p=0.407), as seen in Table 5.14.

Table 5.14 Number of studies based on the number of researchers with and without multiple testing correction

Number of researchers	Number of studies (n=139) (%)	Multiple testing correction (n=40) (%)	No multiple testing correction (n=99) (%)	p value
1-4	75 (54.0%)	25 (33.3%)	50 (66.7%)	0.407
5-7	58 (41.7%)	14 (24.1%)	44 (75.9%)	
8 or more	6 (4.3%)	1 (16.7%)	5 (83.3%)	

5.6.5 Statistician / epidemiologist involvement

The involvement of a statistician/ epidemiologist was not associated with multiple testing correction (p=0.295). The assessment of whether a statistician/ epidemiologist was involved in the published articles was determined by the affiliation information for the authors. When the exact position of the authors was unclear, the name and university was further investigated online via Google to determine whether if they were a statistician involved in the study.

The majority of the papers (94.2%, n=131) with multiple testing did not involve a statistician/ epidemiologist. There was only a total of 8 publications (5.8%) which had statistician involvement, however only 1 paper that had accounted for multiple testing (12.5%), as illustrated in Table 5.15.

Table 5.15 Number of studies based on the statistician/ epidemiologist involvement with and without multiple testing correction

Statistician/ epidemiologist involvement	**Number of studies (n=139) (%)**	**Multiple testing correction (n=40) (%)**	**No multiple testing correction (n=99) (%)**	**p value**
Yes	8 (5.8%)	1 (12.5%)	7 (87.5%)	0.295
No	131 (94.2%)	39 (29.8%)	92 (70.2%)	

Table 5.16 Distribution of 139 articles with multiple hypothesis testing based on study type, journal classification, region of authorship, number of researchers and statistician/ epidemiologist involvement

Variables	Number of studies (n=139) (%)		Multiple Testing Correction (n=40) (%)	No Multiple Testing Correction (n=99) (%)	p value
Study type	Prospective	53 (38.1%)	18 (34.0%)	35 (66.0%)	0.289
	Retrospective	86 (61.9%)	22 (25.6%)	64 (74.4%)	
Journal classification	Main orthodontic journal	57 (41%)	25 (43.9%)	32 (56.1%)	**0.002***
	Non main orthodontic journal	82 (59%)	15 (18.3%)	67 (81.7%)	
Region of authorship	Americas	28 (20.1%)	5 (17.9%)	23 (82.1%)	0.093
	Europe	33 (23.7%)	14 (42.4%)	19 (57.6%)	
	Asia/ Other	78 56.2%)	21 (26.9%)	57 (73.1%)	
Number of researchers	1-4	75 (54.0%)	25 (33.3%)	50 (66.7%)	0.407
	5-7	58 (41.7%)	14 (24.1%)	44 (75.9%)	
	8 or more	6 (4.3%)	1 (16.7%)	5 (83.3%)	
Statistician / epidemiologist involvement	Yes	8 (5.8%)	1 (12.5%)	7 (87.5%)	0.295
	No	131 (94.2%)	39 (29.8%)	92 (70.2%)	

* The p value would still remain significant after application of the Bonferroni correction, considering 5 statistical tests were undertaken to examine factors associated with multiple testing correction

5.7 Inter and intra-reliability testing

5.7.1 Title and abstract screening

The kappa score for intra- and inter-examiner reliability were 1.0 and 0.935 indicating excellent intra- and inter-examiner reliability during title and abstract screening.

5.7.2 Data extraction

Intra-rater reliability was not assessed using percentage agreement and Kappa statistics. When the data was compared on the ten percent random sample of the included papers, majority of the extracted data were similar with only minor technical error detected. Therefore, it was decided that no formal statistical analysis to be carried out to assess intra-rater reliability in view of the nature and amount of the extracted data which may complicate the calculation of the kappa score.

Chapter 6: Discussion

6.1 Summary of the overall findings

The total number of published studies associated with multiple testing in relation to the use of lateral cephalometric variables as the outcome measure was 139. This resulted from both electronic searching and handsearching for studies published between 1st January 2014 to 31st December 2015.

The reported mean number of variables in a cephalometric study was 20. If each of the cephalometric variables was subjected to one hypothesis test (provided all were independent of each other), this would result in a total of 20 hypothesis tests. Moreover, there was an average of 3 families of tests (inter/ intra-group comparison and comparison across different time points) using a list of lateral cephalometric variables in each study. When exploring in detail, the use of lateral cephalometric variables essentially involves several levels of multiplicity problem (e.g. comparison of more than two groups, repeated measurements of each endpoint and comparison of multiple outcome measurements). However, this review aimed to only examine the multiplicity problem particularly with the use of the lateral cephalometric variables as the outcome measure.

This study found only 40 studies (29%) in some way corrected or accounted for multiple testing. This reflected that the correction methods are not widely applied in orthodontic literature. Most of the results were interpreted solely based on the p values from significance testing to draw conclusions on the effectiveness of a treatment. Pandis pointed out that interpretation of p value from multiple comparisons might be misleading as it does not provide sufficient information about the effect size of a treatment but rather p value on its own only provides the strength of the evidence against the null hypothesis.[40]

Most authors suggested the use of confidence intervals (CIs) as another alternative of reporting statistical significance.[28,37,38,42] Confidence interval provides information on the range of the effect point estimate and allows quantification of the precision of the results. It is usually set at 95 per cent, in which we are 95 per cent confident that the true population value lies.[28,42] The midpoint of the interval which is the point estimate is the indication of the range of the difference of effect between the groups. The precision of the results is determined by the width (range) of the CI and it is sensitive to the standard deviation (SD) and sample size.[28]

Increasing the sample size will narrow the width of the CIs around the similar size of difference, therefore it increases precision. A narrow CI shows good precision whereas a wide

CI shows a poor precision that the result should be viewed with caution as the certainties in determining range of effect size is questionable.[28,42]

On the other hand, as in the case of p value where by increasing the sample size, it lowers the p value. Therefore, CIs reporting shifts the interpretation of results from either a significant or non-significant approach to the size of the effect and its range which offer valuable information when making a clinical decision.[42]

Various methods were used to account for multiple testing. 21 publications (52.5%) that have accounted for multiple testing had a pre-specified primary outcome that was adhered to when making conclusions of the study. This was followed by ten publications with statistical correction (25%), five which claimed to be preliminary studies (12.5%), two which were pilot studies (5%), one stated as an exploratory study (2.5%) and one study aimed at generating hypotheses (2.5%). Hence, it could be postulated that majority of the orthodontic studies using cephalometric variables as the outcome measure neither consider statistical correction nor have a pre-specified primary outcome from the outset.

When examining the rationales for multiple testing correction in the included studies, only a minority of the studies (n=6) stated the rationale of the statistical correction. Of the studies that provided a discussion of the correction, the main consideration was its relevance in controlling type I error in multiple testing. This again showed a lack of understanding in recognising multiplicity problem with the use of lateral cephalometric variables as the outcome measure. This is imperative as the multiplicity issues can be addressed during the study design stage and also the analytical phase of a research.

As mentioned earlier, the majority of papers did not consider the relative risk of type I error and this was reflected in the error rates calculated from the included studies. The chance of obtaining at least one false positive (type I error) was estimated at 58%, with an average of 13% of the p values labelled as significant possibly arising just by chance alone. Interestingly, when a simple correction experiment using the Bonferroni correction by amalgamating all the significant p values from the included studies was carried out, it resulted in only 47% of the p values that remained significant. In essence, it is important that the researchers are aware of the effects of multiple testing when using multiple lateral cephalometric variables as the outcome measure in a cephalometric study.

Factors influencing whether a study accounted for multiple testing were also examined. The included papers were investigated with regards to various trial characteristics which included the following variables:

- Study type
- Journal classification
- Region of authorship
- Number of researchers
- Statistician/ epidemiologist involvement

The trial characteristics were chosen once all the included papers had been examined as these characteristics varied significantly between different study reports. Of the variables examined, the only significant association with the application of multiple hypothesis testing correction was the journal classification ($p=0.002$). Fundamentally, this reflected that accurate statistical reporting and the use of appropriate correction to account for multiple testing are fundamental for studies to be published in main orthodontic journals with high SJR factor.

6.2 Combination of electronic searching and handsearching

All papers were electronically searched on four electronic databases (PubMed, Ovid Medline, Scopus and EBSCO Dentistry & Oral Sciences Source) and handsearched within four main orthodontic journals (AJODO, AO, EJO and JO) over the two-year period from 1st January 2014 to 31st December 2015 to identify suitable studies for this review. A combination of both electronic searching and handsearching allows amalgamation of all potential studies with multiple testing whether those studies were published in orthodontic or non-orthodontic journals, thus reducing the selection bias. It was pointed out by Mavropoulous and Kiliaridis that approximately half (45%) of the studies were published in non-orthodontic journals with some of the journals having a high impact factor.[26] Therefore, there was a possibility that a potential large amount of information could have been missed by only looking at the publications within the orthodontic journals.

6.3 Electronic searching versus handsearching as the gold standard

Studies have shown that electronic searching has some limitations over handsearching.[102–105] Some of the studies are missed from electronic searching because of the issues with indexing terms, publication in journals not indexed in the main healthcare databases or lack of cover-to-cover indexing.[105]

The shortfall with the use of only electronic database without handsearching was highlighted in the orthodontic literature by Bickley and Harrison. The authors conducted an analysis to compare the search results from both the electronic (MEDLINE) and handsearching. From this simple investigation, only 143 out of 304 studies (47%) were identified by using the electronic database alone.[104] This, in turn emphasises that handsearching of journals is an essential element of a thorough systematic review.

Similarly, a Cochrane Review conducted by Hopewell and co-workers revealed that 92% to 100% of the randomised trials were identified with handsearching alone.[102] The sensitivity (proportion of the total number of known RCTs identified by the search)[106] reduced substantially when electronic searching was performed on different electronic databases with the reported retrieval rate of 49% to 67%.[102] It was noted that variation in the retrieval rate between different electronic databases depended on the complexity of the search strategy.[102] To conclude, a combination of electronic searching and handsearching is the key in achieving a comprehensive search with a high retrieval rate.[102,103]

Comparing to the literature, this study also demonstrated inadequacy of the electronic searching on the retrieval rate of the studies. The number of papers retrieved was underestimated by 10% (n=5) when only using electronic search method alone. The retrieval rate was indeed high when compared to the findings from Cochrane Review.[102] This may be explained by the fact that the four orthodontic journals (AJODO, AO, EJO and JO) were indexed in the chosen electronic databases in this study. However, if the sensitivity of the electronic searching is to be improved, the search terms may need to be devised using a combination of controlled vocabulary and free-text terms which suits each electronic database if more papers are not to be missed in the future.

6.4 Selection of orthodontic journals for handsearching

Sun et al. identified that 45% of the articles from 1990 to 1998 were published in five orthodontic journals that provided clinically relevant information to orthodontists. The five key journals were the AJODO, AO, EJO, JO and International Journal of Adult Orthodontics and Orthognathic Surgery (ceased publication in 2002).[107] Furthermore, the selection of the four orthodontic journals namely AJODO, AO, EJO and JO were also recommended by Shimada et al. for practice of evidence-based orthodontics in order to gather high quality materials related to orthodontics.[108] Hence, these four orthodontic journals were selected for handsearching in this study. Since the International Journal of Adult Orthodontics and Orthognathic Surgery terminated its publication in 2002, this journal was excluded from this review.

6.5 Classification of journals based on SCImago Journal and Country Rank (SJR)

During the data analysis stage, it was considered that journal classification should be based on journal metrics that evaluate the general citation impact of the journals. It allows measurement of a journal's impact and its quality relative to the other journals in the respective subject field. Hence, the journal classification into main orthodontic journals or non-main orthodontic journals according to the journal metrics would appear to be more compelling. The classification of the included articles (main and non-main orthodontic journals) was then used to examine whether this was a factor influencing multiple hypothesis correction.

There are numerous bibliometric indicators with the Web of Science and Scopus that form the two most dominant citation indexing databases evaluating and ranking the journals through indices including Impact Factor (IF) which is published in the Journal of Citation Report (JCR)[109] and SCImago Journal and Country Rank (SJR).[100]

The SJR indicator is calculated from the data held in the Scopus citation database. It is a prestige metric taking into account the subject areas, quality and reputation of the journal that influences the value of the citation. Unlike the traditional citation method that considers all citations to be 'equal', SJR indicator measures the scientific impact of a journal from two distinct perspectives which are the number of citations received by the journal and also the prestige of the journals being cited. Additionally, SJR offers a good control to prevent journal manipulation through the use of self-citation by limiting the self-citation rate to a maximum of 33%. SJR indicator is computed from the mean number of weighted citations received in the selected year by the number of articles published in the selected journal in the previous three-year citation period.[110]

Journal of Citation Report (JCR) based on the Impact Factor (IF) is the most commonly used indices in assessing the journals' scientific impact and quality which was first introduced by Garfield in 1955. IF is computed based on the ratio of the recorded number of citations within a particular year to the published items during the two preceding years, divided by the total number of articles in the same two years.[111]

In this review, SJR indicator was selected rather than the most well-known and commonly used IF as a tool for journal classification. Firstly, the number of the indexed journals in the subject area 'Dentistry' is substantially higher in Scopus[100] database than the Web of Science.[109] In 2015, the Scopus database comprised of 178 dental journals in its directory and only 91 dental journals in the Web of Science database.[100,109] On top of that, SJR allows grouping of journals based on subject categories 'Orthodontics' which appears to be relevant

to this study.[100] In contrast, JCR contains broader subject category which is 'Dentistry, Oral Surgery & Medicine'.[109] Furthermore, SJR indicator has demonstrated its high correlation with the IF across a number of different disciplines.[112–115] A recent study has shown that SJR indicator is highly correlated with IF that it offers another alternative for researchers in the dental field for assessment of the quality of the research.[116] Another merit of the SJR indicator is its open access resource from the internet while JCR is a commercial product that requires subscription.[116]

The top four ranking journals based on 2015 SJR indicator were classified as main orthodontic journals (see Table 6.1).[100] SJR Indicator in 2015 was used instead because 2016 data was yet to be published during the data analysis stage.

Table 6.1 Top four orthodontic journals in 2015 based on SJR indicator

No.	Name of Journal	SJR Indicator
1.	American Journal of Orthodontics and Dentofacial Orthopedics (AJODO)	1.343
2.	Angle Orthodontist (AO)	1.313
3.	European Journal of Orthodontics (EJO)	1.129
4.	Korean Journal of Orthodontics (KJO)	0.982

6.6 Comparison of findings with previous published research

6.6.1 Summary of the studies with multiple hypothesis testing

No studies in the field of dentistry were identified which examined articles with multiple hypothesis testing. Therefore, the findings were compared to the previous similar studies in the medical literature. A summary of the studies looking at multiple testing in different medical specialties is illustrated in Table 6.2.

The present study had a similar approach to the review by McClean and Silverberg[12] which carried out a combination of electronic and handsearching of the journals to identify papers with multiple testing. Otherwise, the majority of the publications only performed handsearching of the journals in their respective medical area.[3,6,8,9,36]

Armstrong reported the highest proportion of studies (67%) that addressed the problem of multiple testing by the Bonferroni correction in Ophthalmic research when comparing to other

similar studies.[36] A possible explanation was the inclusion of articles with any correction methods including a Bonferroni post-hoc analysis following ANOVA.

In contrast, the present investigation showed a slightly higher percentage of papers in some way have corrected or accounted for multiple testing (29%) when compared to other studies. In this review, if a study conclusion was made based on the pre-specified primary outcome measure, the study would be considered that it has in some way addressed the multiplicity problem. However, it should be noted that the interpretation made in this context may have biased the findings as there was inherent subjectivity in interpreting the study conclusion and also the examiner (SCP) was not blinded to the information on all the included articles. In addition, if a study was exploratory in nature, the included articles were also considered to have accounted for multiple testing.

Most of the publications examined the full-text articles whereas one study only looked at the abstracts of the studies. Stacey et al. examined the abstracts of the publications presented at The Association for Research in Vision and Ophthalmology (ARVO) in May 2010 with only 3.2% of the abstracts mentioning some form of corrections for multiple testing which was the lowest among all the other studies.[9] This was likely to be an underestimate as limited information could be extracted from the conference abstracts. In addition, there is a possibility that an abstract may not contain any mention of a multiple comparisons correction, however the actual paper does.

Ottenbacher aimed to only examine type 1 error rates in a sample of published studies with random selection of five issues of journals published in 1996 in the American Journal of Public Health and the American Journal of Epidemiology.[8] As a result, quantification of the articles associated with multiple testing and the proportion of papers that have accounted for multiple testing was impossible as the total number of articles published in 1996 in both journals were not rigorously searched.

Bonferroni correction was the most commonly applied statistical correction, as can be seen in Table 6.2. This was likely due to its ease of application when performing statistical correction for multiple testing. However, this comes at the expense of reducing the study power and that this may not be optimal for studies with smaller sample size which had been previously discussed in the literature review.

Table 6.2 Summary of studies looking at multiple hypothesis testing

Studies	Medical/ dental field	Number of studies	Study period	Search method	
				Handsearching	Electronic searching
McClean and Silverberg 2015[12]	Dermatology	162	1 May 2013-1 May 2014	Electronic searching on MEDLINE and handsearching within 44 dermatology journals with studies limited to RCTs	
Kirkham and Weaver 2015[6]*	Otolaryngology	140	2012	Four journals-The Laryngoscope, Archives of Otolaryngology-Head & Neck Surgery (now known as JAMA Otolaryngology-Head & Neck Surgery), Otolaryngology-Head & Neck Surgery and Annals of Otology, Rhinology and Laryngology	-
Armstrong 2014[36]	Ophthalmology	142	2003-2013	Three journals-Ophthalmic & Physiological Optics (OPO), Optometry & Vision Sciences (OVS) and Clinical & Experimental Optometry (CXO)	-
Walenkamp et al. 2013[3]*	Orthopedic	127	2010	Two journals-Journal of Bone and Joint Surgery American Edition and Journal of Bone and Joint Surgery British Edition	-
Stacey et al. 2012[9]	Ophthalmology	2321 abstracts	2010	All abstracts from presentations at The Association for Research in Vision and Ophthalmology (ARVO)	-
Ottenbacher 1998[8]	Public Health and Epidemiology	173	1996	Two journals-American Journal of Public Health & American Journal of Epidemiology	-
Current study*	Orthodontics	139	2014-2015	Combination of electronic searching on four electronic databases-PubMed, Ovid Medline, Scopus and EBSCO Dentistry & Oral Sciences Source and handsearching in four orthodontic journals-AJODO, AO, EJO and JO	

Studies	Number of studies with multiple testing correction	Methods of correction (in some way corrected or accounted for multiple testing)	Most commonly used statistical correction	
McClean and Silverberg 2015[12]	16 (9.9%)	Bonferroni, Holm-Bonferroni's and Dunn's	Bonferroni	7 (43.8%)
Kirkham and Weaver 2015[6]*	14 (10%) including 8 with statistical correction	Statistical correction with Bonferroni, Tukey Kramer method, False Discovery Rate and a decreased significance level of 0.005 without citing method of correction Others: Independent validation of results, pre-specified primary versus secondary outcomes, study of exploratory in nature, mentioned about multiple testing and/ or the need for further validation	Bonferroni	5 (62.5%)
Armstrong 2014[36]	95 (67%)	Bonferroni, Bonferroni-Holm, standard Abbott formula, False Discovery Rate, Hochberg method, or alternative post-hoc procedure such as Scheffe's test	Bonferroni	51 (36%)
Walenkamp et al. 2013[3]*	14 (11%)	Bonferroni and a decreased significance level of 0.01 Others: mentioned a correction but method was not described	Bonferroni	12 (86%)
Stacey et al. 2012[9]	74 (3.2%)	Bonferroni, Tukey's, False Discovery Rate, Least Significant Difference, Dunnett's, Scheffe's, Newman Keul's and non-specific multiple comparison test	Bonferroni	24 (32%)
Ottenbacher 1998[8]	-	-	-	-
Current study*	40 (29%) including 10 with statistical correction	Statistical correction with Bonferroni, Benjamini Hochberg & a decreased significance level of 0.01 Others: Primary outcome stated and adhered to when making conclusions, studies which were preliminary, exploratory, hypothesis driven and pilot study	Bonferroni	7 (70%)

* Multiple testing was defined as testing five or more hypotheses within a family of inferences.

6.6.2 Error rates for studies with unaccounted multiple testing

When comparing the error rates for studies with unaccounted multiple hypothesis testing (see Table 6.3), the results of this review were relatively comparable to the findings from the medical literature.[3,6,8] The similarity in the reported error rates highlighted that the inflation of type I error is common in both medical and dental specialties.

In the present study, the mean probability of making at least one type I error in a family of inferences (family-wise error rate) was 0.58 with the value appearing to be the average among the other three studies. This means that at least one test will be significant (if all null hypotheses are true) is 58% which is nearly 12 times the original alpha (α) level of 0.05.

The mean expected number of type I error in a particular group of statistical significance tests (error rate per experiment) was marginally higher at 0.97. Therefore, at the alpha (α) level of 0.05, one would expect one type I error in a family of inferences, which indirectly suggested that majority of the cephalometric studies involved a list of 20 cephalometric variables. However, the interpretation of error rate per experiment is not straightforward because there is no upper limit for its value.

The mean percentage of results labelled as statistically significant that were likely to be by chance alone (percent error rate) was marginally lower at 13.44%. At 0.05 significance level, 5% is the lowest bound value for the percent error rate. For example, if 1 out of 20 comparisons evaluated at the 0.05 significance level is statistically significant, the percent error rate is 100% (the formula of the percent error rate has been discussed in the literature review section 2.2.1.3), indicating that the number of tests found to be significant, which is 1, is the number expected by chance. On the other hand, if 5 out of 20 comparisons are significant, the percent error rate is 20%, suggesting that 20% of the results are expected by chance (1 test) and the remaining 80% (4 tests) are likely to be caused by non-chance factors.

Table 6.3 Comparison of the error rates between studies for articles with unaccounted multiple hypothesis testing (mean± SD)

Studies	Family-wise error rate	Error rate per experiment	Error rate (%)
Kirkham and Weaver 2015[6]	0.41± 0.17	0.61± 0.78	18± 29
Ottenbacher 1998[8]	0.68± 0.24[a]	0.90± 0.57	19.16± 9.01
	0.70± 0.29[b]	0.87± 0.51	18.73± 9.32
Walenkamp et al. 2013[3]*	54% (34-81)[c]	-	-
	54% (34-66)[d]		
Current study	0.58± 0.19	0.97± 0.51	13.44± 11.93

* Error rate presented in median and interquartile range (%)

a- American Journal of Public Health

b- American Journal of Epidemiology

c- Journal of Bone and Joint Surgery American Edition

d- Journal of Bone and Joint Surgery British Edition

6.6.3 Comparison of studies with reported number of hypothesis tests

The number of hypothesis tests in both the studies (McClean and Silverberg[12] and present study) was similar with a reported mean of 20.9 and 19.63 hypothesis tests, as outlined in Table 6.4. This suggested that as the number of hypothesis tests increases, the probability of making at least one type I error also escalates proportionally. For reference, the family-wise error rate when performing 20 statistical tests is 64% which means that a cephalometric study with 20 cephalometric variables has a 64% probability of falsely rejecting the null hypothesis (false positive finding).

Table 6.4 Comparison of studies with reported number of hypothesis tests

Studies	Hypothesis tests (per family of tests)	
	Mean ± SD	Range
McClean and Silverberg 2015[12]	20.9 ± 19.2	2-108
Current study	19.63 ± 9.79	5-47

6.6.4 The correction experiment

The Bonferroni correction was applied to the significant p values within each family of inferences to yield the corrected results on the studies that have not accounted for multiple testing. The Bonferroni's method was chosen as it is one of the classical corrections for p value adjustment with its ease of application.

With the application of the Bonferroni method, as seen in Table 6.5, both studies (Kirkham and Weaver[6] and present study) showed a substantial reduction in the p values that remained significant. In the current study, the reduction of significant p values was slightly higher at 53% when compared to the study by Kirkham and Weaver[6] with 43% reduction of the significant p values. This reflected that, on average, multiple testing that did not account for correction would likely to inflate the statistical significant findings by 50%. Hence, the multiplicity issue has some implications when it comes to interpreting the significant cephalometric findings e.g. assessment of the effectiveness of an orthodontic appliance because a proportion of the significant results might merely be just the false positive findings.

Table 6.5 Comparison of findings on the correction experiment

Studies	**p value < 0.05**	**p value that remained significant after Bonferroni correction**	**Reduction of significant p value (%)**
Kirkham and Weaver 2015[6]	1509	860	43%
Current study	2385	1115	53%

6.7 Limitations of the study

6.7.1 Design of the study

To date, there were no previous publications in the orthodontic literature that examined the extent of the multiple testing in relation to the use of multiple cephalometric variables as the outcome measure. This caused difficulty in designing a study to quantify multiple hypothesis testing within the recent published lateral cephalometric studies in orthodontics because there were no similar studies from any published orthodontic literature to be used as a reference from the outset. Hence, the foundation of this study was mainly based on the previous published medical literature looking at multiple testing with further refinement made to meet the aim and objectives of this study.

6.7.2 Inclusion and exclusion criteria

In order to only include studies with multiple testing that were relevant to orthodontics, this study had a stringent inclusion and exclusion criterion. The cephalometric assessment in the dento-facial region is of particular interest to orthodontist to assess dento-skeletal proportions and to determine the underlying aetiology of the dental malocclusion for diagnosis and treatment planning. Therefore, any forms of cephalometric evaluation other than from the dento-facial region (e.g. airway and cranial base) were excluded.

In addition, it is possible to reconstruct or generate lateral cephalometric images from CBCT using appropriate software to produce volume rendered or surface rendered images. This is known as the synthesised cephalometric images. A large field of view (FOV) is therefore required to obtain a rendered cephalometric image from CBCT. However, there is no indication to practice large FOV CBCT with a view to obtain cephalometric data.[69,117] Therefore, it is thought that limiting the studies to only include 2-dimensional lateral cephalometric radiograph in this study is appropriate as this imaging technique is relatively applicable to daily orthodontic practice.

6.7.3 Identification of papers

This was a retrospective, observational study that fundamentally was open to bias. A robust search for every single publication with multiple testing using numerous cephalometric variables as the outcome measure within each single orthodontic journal was beyond the resources available for this review. Hence, the finding might not represent the overall prevalence of the multiple testing within orthodontic literature.

In addition, there is a possibility of mistakes which were made due to human errors in which papers which should have been included were unintentionally omitted. However, precautions were taken at title and abstract screening and data extraction stage by examining a 5% random sample of the included studies which demonstrated good inter and intra-examiner reliability. As electronic searching was carried out on a number of electronic databases, it was thought that an additional handsearching would be beneficial as it may help in locating additional published studies which could have been missed from electronic searching. As a result of the handsearching from the four orthodontic journals, five additional papers (10%) were found.

The choice of only selecting studies associated with multiple testing with a minimum of five or more hypothesis tests would have resulted in a lower percentage of multiple testing articles within those reviewed. This was proven in a study which assessed the reporting characteristics of the articles published in six major clinical dental specialty journals. The authors found that multiple comparisons were relatively prevalent with an overall findings of 42.7%, 31.5% and 25.8% reporting less than 5, between 5 and 20 and more than 20 comparisons respectively.[118]

However, a higher threshold in defining the number of hypothesis tests within a family of inferences (number of cephalometric variables used as the outcome measure) was aimed to reflect the maximum extent of multiple hypothesis testing within each of the included study, in which this would have resulted in an upper estimates of the number of cephalometric variables used in a cephalometric study which were associated with multiple hypothesis testing.

This study limited the inclusion of the studies to those published in English language only. This inadequacy could have potentially underestimated the studies associated with multiple testing due to language bias, which may have an impact on the findings of the study.

6.7.4 Data extraction

There was no effort made to contact the authors of the included studies for study clarification especially on the statistical methods as this was beyond the remit of this study. Hence, it was clearly stated from the outset of the study that if any of the included studies has insufficient information or unclear statistical methods, the study would be omitted.

In view of the large volume of data to be extracted from the included papers, main consideration was placed on the amount of time needed for completion of the data extraction and data analysis. The review was therefore limited to a two-year period publication rather than looking at the initial study sample over a three-year period. When comparing to the other

similar studies in the medical literature, most of the studies looked at one to two-year period of publication with inclusion of a total number of 127-173 articles that were associated with multiple testing (See Table 6.2).[3,6,8,9,12] Therefore, the decision to restrict the review to a two-year period was deemed to be reasonable. Even though this study was limited to a two-year period, it was considered to be comparable to other previous published studies because a total of 139 publications were included in this study.

6.7.5 Data analysis

The assumption made on the independence of the tests may not be true for all papers that have been included in the data analysis especially in the calculation of the error-rates. As an example, the cephalometric variables used in the study may be correlated because the linear and angular measurements are taken from the intersection of reference planes based on the anatomic landmarks on the lateral cephalogram in the dento-facial region. The cephalometric variables used therefore may be correlated and not independent of one another. The error rates are lower if the tests are correlated. Therefore, the error rates obtained should be considered as an 'upper limit' of the actual values in this review.

6.7.6 Quality

During the data collection stage, there were no attempts made to assess the methodological quality of the included studies. It was thought to be outside the remit of this study and this should be analysed in the future if similar study is to be undertaken. Methodological quality assessment is relevant if the reported results of the studies are being assessed. This study aimed to look at the issue of multiple hypothesis testing in relation to the use of lateral cephalometric variables as the outcome measure, which arguably is one of the aspects of methodological quality assessment.

6.7.7 Reliability

For title and abstract screening, the kappa score for intra- and inter-examiner reliability were 1.0 and 0.935 indicating excellent intra- and inter-examiner reliability. This highlighted that the title and abstract screening was systematically screened independently by SCP with a level of good consistency in order to avoid inadvertent omission of any relevant papers. However, intra-rater reliability was not assessed during data extraction and this was discussed in the result section 5.7.2.

6.8 Research implications

This review shed some lights onto the multiple testing problem in relation to the use of multiple lateral cephalometric variables as the outcome measure in the orthodontic literature. Additionally, the corrections for multiple testing are not widely applied. The potential inflation of type I error as illustrated was concerning that necessary precautions should be taken. As suggested in the previous published literature, the following key recommendations have been made for both authors and readers.

6.8.1 Authors strategies:

The following suggested strategies are for the authors to consider when faced with multiple comparisons using numerous cephalometric variables:

1) Pre-specify a primary outcome at the study design stage[3,16,28,119–121]
2) Acknowledge the potential type I and type II errors with its possible consequences to the reader[120]
3) Use of a composite endpoint limited to the area of interest e.g. maxilla, mandible and dentition[27]
4) Limit the number of cephalometric variables in a study[28–32]
5) Pre-define the secondary outcomes (if any) to avoid 'fishing expedition[3]
6) Perform formal and appropriate statistical correction for confirmatory study [3,7,9,119,121,122] even though there is no gold standard for multiple test adjustment as yet
7) Use of significance testing of multivariate methods e.g. MANOVA or Hotelling's T^2 test, global test statistics developed by O'Brien[123] and further modified by Pocock et al.[124] and Exact tests developed by Lauter[125]

6.8.2 Readers strategies:

The following approaches should allow the readers to reach a sensible conclusion from study involving multiple testing, irrespective of whether the articles consider any methods of multiple testing correction.

1) Have a good understanding on complicated study design by evaluating the study quality and be well equipped with a good knowledge in statistics in order to evaluate the effect size of the findings prior to interpreting the statistical significance[33]
2) Beware of 'data dredging' or 'p value hunting' from multiple testing in relation to the use of numerous cephalometric variables[120]
3) Seek information in the methodology section of the articles whether there is a pre-specified primary outcome measure and interpret any additional findings in this context[44]

6.9 Direction for future research

The following recommendations have been made for future research.

1) To repeat in approximately another decade to assess the extent of the multiple testing problem in the relation to the use of lateral cephalometric variables as the outcome measure.
2) To assess the methodological quality of the included articles to improve the robustness of the review.
3) To explore the relationship between the inter-related lateral cephalometric variables because the comparison of one cephalometric variable may be confounded by its relationship with another cephalometric variable. Therefore, p value adjustment should take into account the correlation among the cephalometric variables.
4) To explore the search in articles published in non-English language to reduce the risk of language bias.

Chapter 7: Conclusions

1) Multiple testing is common in orthodontic research especially in relation to the use of multiple cephalometric variables as the outcome measurement.
2) A total of 139 studies published over a two-year period were included in this review with only 29% of the articles (n=40) considered the effect of multiple testing that these studies in some way have corrected or accounted for multiple testing.
3) The potential prevalence of false positive results in the sample of published articles in orthodontic research in relation to the use of lateral cephalometric variables were (mean± SD):
    - family-wise error rate: 0.58 ± 0.19
    - error rate per experiment: 0.97 ± 0.51
    - percent error rate: 13.44 ± 11.93
4) Of the 40 articles that addressed the problem of multiple hypothesis testing, ten applied statistical correction (25%), twenty-one pre-specified a primary outcome which was adhered to when making conclusions from the study (52.5%), five claimed to be preliminary studies (12.5%), two were pilot studies (5%), one stated as an exploratory study (2.5%) and one study aimed at generating hypotheses (2.5%).
5) Of the ten applying statistical correction, Bonferroni correction is the most commonly used method (70%), followed by Benjamini Hochberg (20%) and one chose a decreased significance level of 0.01 without specifying any correction method (10%).
6) The only statistically significant factor influencing the application of the multiple testing correction was the journal classification with journals published in the main orthodontic journals were more likely to account for multiple hypothesis testing (p=0.002).
7) Handsearching was superior than electronic searching with 10% of papers (n=5) that were missed from electronic searching.

Chapter 8: References

1. Petrie A, Sabin C. Hypothesis testing. In: *Medical Statistics at a Glance*. 3rd ed. Wiley-Blackwell; 2009:50-53.
2. Daniel WW, Cross CL. Hypothesis testing. In: *Biostatistics-A Foundation for Analysis in the Health Sciences*. 10th editi. Wiley; 2013:214-303.
3. Walenkamp MM, Roes KC, Bhandari M, Goslings JC, Schep NW. Multiple testing in orthopedic literature: A common problem? *BMC Res Notes*. 2013;6:374.
4. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. *Ind Psychiatry J*. 2009;18(2):127-131.
5. Gordi T, Khamis H. Simple solution to a common statistical problem: Interpreting multiple tests. *Clin Ther*. 2004;26(5):780-786.
6. Kirkham EM, Weaver EM. A review of multiple hypothesis testing in otolaryngology literature. *Laryngoscope*. 2015;125(3):599-603.
7. Sainani KL. The problem of multiple testing. *Am Acad Phys Med Rehabil*. 2009;1(12):1098-1103.
8. Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. *Am J Epidemiol*. 1998;147(7):615-619.
9. Stacey AW, Pouly S, Czyz CN. An analysis of the use of multiple comparison corrections in Ophthalmology research. *Invest Ophthalmol Vis Sci*. 2012;53(4):1830-1834.
10. Tyler KM, Normand S-LT, Horton NJ. The use and abuse of multiple outcomes in randomized controlled depression trials. *Contemp Clin Trials*. 2011;32(2):299-304.
11. Vickerstaff V, Ambler G, King M, Nazareth I, Omar RZ. Are multiple primary outcomes analysed appropriately in randomised controlled trials? A review. *Contemp Clin Trials*. 2015;45(Part A):8-12.
12. McClean M, Silverberg JI. Statistical reporting in randomized controlled trials from the dermatology literature: A review of 44 dermatology journals. *Br J Dermatol*. 2015;173(1):172-183.
13. Abdi H. The Bonferonni and Sidak corrections for multiple comparisons. In: *Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA): Sage; 2007.
14. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm Methods. *Am J Public Health*. 1996;86(5):726-728.
15. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65-70.
16. Bender R, Lange S. Adjusting for multiple testing- When and how? *J Clin Epidemiol*. 2001;54(4):343-349.

17. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *J R Stat Soc*. 1995;57(1):289-300.
18. Godfrey K. Statistics in practice- Comparing the means of several groups. *New Engl J Med*. 1985;313(23):1450-1456.
19. Streiner DL. Commentary #11- Multiple comparisons and peeking at data. *J Clin Psychopharmacol*. 2016;36(1):5-8.
20. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J Clin Epidemiol*. 2014;67(8):850-857.
21. Simas R, Maestri F, Normando D. Controlling false positive rates in research and its clinical implications. *Dental Press J Orthod*. 2014;19(3):24-25.
22. Baumgartner S, Pandis N, Eliades T. Exploring the publications in three major orthodontic journals- A comparative analysis of two 5-year periods. *Angle Orthod*. 2014;84(3):397-403.
23. Wahl N. Orthodontics in 3 millennia. Chapter 8: The cephalometer takes its place in the orthodontic armamentarium. *Am J Orthod Dentofac Orthop*. 2006;129(4):574-580.
24. Koletsi D, Fleming PS, Eliades T, Pandis N. The evidence from systematic reviews and meta-analyses published in orthodontic literature. Where do we stand? *Eur J Orthod*. 2015;37(6):603-609.
25. Gibson R, Harrison J. What are we reading? An analysis of the orthodontic literature 1999 to 2008. *Am J Orthod Dentofac Orthop*. 2011;139(5):e471-e484.
26. Mavropoulos A, Kiliaridis S. Orthodontic literature: An overview of the last 2 decades. *Am J Orthod Dentofac Orthop*. 2003;124(1):30-40.
27. Pandis N. Multiplicity 2: Multiple treatments and multiple outcomes. *Am J Orthod Dentofac Orthop*. 2013;143(4):589-591.
28. Harrison JE. Evidence-based Orthodontics- How do I assess the evidence? *J Orthod*. 2000;27(2):189-197.
29. O'Brien K, Wright J, Conboy F, et al. Early treatment for Class II Division 1 malocclusion with the Twin-block appliance: A multi-center, randomized, controlled trial. *Am J Orthod Dentofac Orthop*. 2009;135(5):573-579.
30. O'Brien K, Wright J, Conboy F, et al. Effectiveness of early orthodontic treatment with the Twin-block appliance: A multicenter, randomized, controlled trial. Part 1: Dental and skeletal effects. *Am J Orthod Dentofac Orthop*. 2003;124(3):234-243.
31. Tulloch JFC, Proffit WR, Phillips C. Outcomes in a 2-phase randomized clinical trial of early class II treatment. *Am J Orthod Dentofac Orthop*. 2004;125(6):657-667.
32. Tulloch JFC, Phillips C, Koch G, Proffit WR. The effect of early intervention on

skeletal pattern in Class II malocclusion: A randomized clinical trial. *Am J Orthod Dentofac Orthop*. 1997;111(4):391-400.

33. Rinchuse DJ, Close JM, Rinchuse DL, Law SV, Rinchuse DN. Summary of statistics in American Journal of Orthodontics and Dentofacial Orthopedics articles published in 2003. *Am J Orthod Dentofac Orthop*. 2006;130(4):511-515.
34. Law SV, Chudasama DN, Rinchuse DJ. Evidence-based orthodontics Current statistical trends in published articles in one journal. *Angle Orthod*. 2010;80(5):952-956.
35. Whitley E, Ball J. Statistics review 3: Hypothesis testing and P values. *Crit Care*. 2002;6(3):222-225.
36. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt J Coll Optom*. 2014;34:502-508.
37. Rothman KJ. A show of confidence. *New Engl J Med*. 1978;299(24):1362-1363.
38. Nuzzo R. Statistical Errors. *Nature*. 2014;506:150-152.
39. Mainland D. Statistical ritual in clinical journals: Is there a cure? *Br Med J*. 1984;288:841-843.
40. Pandis N. The P value problem. *Am J Orthod Dentofac Orthop*. 2013;143(1):150-151.
41. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, process and purpose. *Am Stat*. 2016;70(2):129-133.
42. Polychronopoulou A, Pandis N, Eliades T. Appropriateness of reporting statistical results in orthodontics: the dominance of P values over confidence intervals. *Eur J Orthod*. 2011;33(1):22-25.
43. Dar R, Serlin RC, Omer H. Misuse of statistical tests in three decades of psychotherapy research. *J Consult Clin Psychol*. 1994;62(1):75-82.
44. Lord SJ, Gebski VJ, Keech AC. Multiple analyses in clinical trials: Sound science or data dredging? *Med J Aust*. 2004;181(8):452-454.
45. Proffit WR, Sarver DM, Ackerman JL. Orthodontic diagnosis: The problem-oriented approach. In: *Contemporary Orthodontics*. 5th ed. Elsevier; 2012:150-219.
46. Sadowsky PL. The geometry of cephalometry. In: *Radiographic Cephalometry-From Basics to Videoimaging*. Quintessence Publishing Co, Inc; 1995:127-136.
47. Cobourne MT, DiBiase AT. The orthodontic patient: examination and diagnosis. In: *Handbook of Orthodontics*. Mosby Elsevier; 2010:125-179.
48. Sterne JAC, Smith GD. Sifting the evidence- what's wrong with significance tests? *BMJ*. 2001;322:226-231.
49. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research- A review of common pitfalls. *Swiss Med Wkly*. 2007;137:44-49.

50. Shaffer JP. Multiple hypothesis testing. *Annu Rev Psychol*. 1995;46(1):561-584.
51. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ*. 1995;310:170.
52. Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc*. 1961;56(293):52-64.
53. Streiner DL, Norman GR. Correction for multiple testing- Is there a resolution? *Chest*. 2011;140(1):16-18.
54. Perneger T V. What's wrong with Bonferroni adjustments. *BMJ*. 1998;316:1236-1238.
55. O'Keefe DJ. Colloquy: Should familywise alpha be adjusted? Against familywise alpha adjustment. *Hum Commun Res*. 2003;29(3):431-447.
56. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1(1):43-46.
57. Kao LS, Green CE. Analysis of Variance: Is there a difference in means and what does It mean? *J Surg Res*. 2008;144(1):158-170.
58. Khamis HJ. Deciding on the correct statistical technique. *J Diagnostic Med Sonogr*. 1992;8:193-198.
59. Mitchell L. Cephalometrics. In: *An Introduction to Orthodontics*. 4th ed. Oxford University Press; 2013:73-84.
60. Allen WI. Historical aspects of roentgenographic cephalometry. *Am J Orthod*. 1963;49(6):451-459.
61. Pacini AJ. Roentgen ray anthropometry of the skull. *J Radiol*. 1922;3:230-238.
62. Broadbent BH. A new x-ray technique and its application to orthodontia. *Angle Orthod*. 1931;1(2):45-66.
63. Hofrath H. Die bedeutung der röntgenfern-und abstandsaufnahme für die diagnostik der kieferanomalien. *Fortschritte der Orthod Theor und Prax*. 1931;1:232-258.
64. Burford D, Newell SL. Cephalometric analysis. In: Gill DS, Naini FB, eds. *Orthodontics: Principles and Practice*. First Edit. Wiley-Blackwell; 2011:78-87.
65. Vitepron S. The technique of cephalometric radiography. In: *Orthodontic Cephalometry*. Mosby-Wolfe; 1995:9-20.
66. Solow B, Tallgren A. Natural head position in standing subjects. *Acta Odontol Scand*. 1971;29(5):591-607.
67. Moorrees CFA. Natural head position- A revival. *Am J Orthod Dentofac Orthop*. 1994;105(5):512-513.
68. Isaacson KG, Thom AR, Atack NE, Horner K, Whaites E. *Guidelines for the Use of Radiographs in Clinical Orthodontics*. 4th ed. British Orthodontic Society; 2015.
69. Abdelkarim AA. Appropriate use of ionizing radiation in orthodontic practice and research. *Am J Orthod Dentofac Orthop*. 2015;147(2):166-168.

70. Broadbent BH. The face of the normal child. *Angle Orthod*. 1937;7(4):183-208.
71. Brodie AG. On the growth pattern of the human head. From the third month to the eighth year of life. *Am J Anat*. 1941;68(2):209-262.
72. Brodie AG, Downs WB, Goldstein A, Myer E. Cephalometric appraisal of orthodontic results- A preliminary report. *Angle Orthod*. 1938;8(4):261-265.
73. Baccetti T, Franchi L, McNamara JA. The Cervical Vertebral Maturation (CVM) method for the assessment of optimal treatment timing in dentofacial orthopedics. *Semin Orthod*. 2005;11(3):119-129.
74. Downs WB. Variations in facial relationship: Their significance in treatment and prognosis. *Angle Orthod*. 1948;34(10):812-840.
75. Steiner CC. Cephalometrics for you and me. *Am J Orthod*. 1953;39(10):729-755.
76. Ricketts RM. A foundation for cephalometric communication. *Am J Orthod*. 1960;46(5):330-357.
77. Jacobson A. The "Wits" appraisal of jaw disharmony. *Am J Orthod*. 1975;67(2):125-138.
78. Mills JRE. The application and importance of cephalometry in orthodontic treatment. *Orthodontist*. 1970;2(2):32-47.
79. McNamara JA. A method of cephalometric evaluation. *Am J Orthod*. 1984;86(6):449-469.
80. Downs WB. Analysis of the dentofacial profile. *Angle Orthod*. 1956;26(4):191-212.
81. Vorhies JM, Adams JW. Polygonic interpretation of cephalometric findings. *Angle Orthod*. 1951;21(4):194-197.
82. Wylie WL. The assessment of anteroposterior dysplasia. *Angle Orthod*. 1947;17(3):97-109.
83. Riedel RA. The relation of maxillary structures to cranium in malocclusion and in normal occlusion. *Angle Orthod*. 1952;22(3):142-145.
84. Steiner CC. Cephalometrics in clinical practice. *Angle Orthod*. 1959;29(1):8-29.
85. Tweed CH. The Frankfort-Mandibular Incisor Angle (FMIA) in orthodontic diagnosis, treatment planning and prognosis. *Angle Orthod*. 1954;24(3):121-169.
86. Sassouni V. A roentgenographic cephalometric analysis of cephalo-facio dental relationships. *Am J Orthod*. 1955;41(10):735-764.
87. Bjork A. *The Face in Profile. An Anthropological X-Ray Investigation on Swedish Children and Conscripts*. Lund: Berlingska Boktryckeriet; 1947.
88. Bjork A. Cephalometric X-ray investigations in dentistry. *Int Dent J*. 1954;4:718-744.
89. Jarabak JR, Fizzell JA. *Technique and Teatment with Light-Wire Edgewise Appliances*. 2nd ed. St Louis: The C. V. Mosby Co.; 1972.

90. Brown M. Eight methods of analysing a cephalogram to establish anteroposterior skeletal discrepancy. *Br J Orthod*. 1981;8(3):139-146.
91. Ballard C. Morphology and treatment of class II dvivision 2 occlusions. *Trans Eur Orthod Soc*. 1956:44-55.
92. Harvold EP. *The Activator in Interceptive Orthodontics*. St Louis: The C.V. Mosby Co.; 1974.
93. Pancherz H. The mechanism of Class II correction in Herbst appliance treatment- A cephalometric investigation. *Am J Orthod*. 1982;82(2):104-113.
94. Holdaway RA. A soft-tissue cephalometric analysis and its use in orthodontic treatment planning. Part I. *Am J Orthod*. 1983;84(1):1-28.
95. Bass NM. The aesthetic analysis of the face. *Eur J Orthod*. 1991;13(5):343-350.
96. Houston WJB. The current status of facial growth prediction: A review. *Br J Orthod*. 1979;6(1):11-17.
97. Viteporn S, Athanasiou AE. Anatomy, radiographic anatomy and cephalometric landmarks of craniofacial skeleton, soft tissue profile, dentition, pharynx and cervical vertebrae. In: *Orthodontic Cephalometry*. Mosby-Wolfe; 1995:21-62.
98. Martins RP, Buschang PH. What is the level of evidence of what you are reading ? *Dental Press J Orthod*. 2015;20(4):22-25.
99. Richardson WS, Wilson MC, Nishikawa J, Hayward RSA. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. 1995;123(3):A12.
100. SCImago Journal & Country Rank. Scopus. http://www.scimagojr.com. Published 2007. Accessed October 18, 2017.
101. GraphPad Software. https://www.graphpad.com/quickcalcs/randomN1.cfm. Published 2017. Accessed September 26, 2016.
102. Hopewell S, Clarke MJ, Lefebvre C, Scherer RW. Handsearching versus electronic searching to identify reports of randomized trials. *Cochrane Database Syst Rev 2007*. (2):Art No.:MR000001.
103. Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials. *Stat Med*. 2002;21(11):1625-1634.
104. Bickley SR, Harrison JE. How to ... find the evidence. *J Orthod*. 2003;30(1):72-78.
105. Cathey JT, Al Hajeri AA, Fedorowicz Z. A comparison of handsearching versus EMBASE searching of the Annals of Saudi Medicine to identify reports of randomized controlled trials. *Ann Saudi Med*. 2006;26(1):49-51.
106. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ*. 1994;309(6964):1286-1291.
107. Sun RL, Conway S, Zawaideh S, Niederman R. Benchmarking the Clinical

Orthodontic Evidence on Medline. *Angle Orthod*. 2000;70(6):464-470.

108. Shimada T, Takayama H, Nakamura Y. Quantity and quality assessment of randomized controlled trials on orthodontic practice in PubMed. *Angle Orthod*. 2010;80(4):713-718.
109. InCites Journal Citation Reports. Clarivate Analytics. https://login.incites.thomsonreuters.com/?DestApp=IC2JCR. Published 2017. Accessed October 18, 2017.
110. Colledge L, de Moya-Anegón F, Guerrero-Bote V, López-Illescas C, El Aisati M, Moed HF. SJR and SNIP: Two new journal metrics in Elsevier's Scopus. *Serials*. 2010;23(3):215-221.
111. Garfield E. The history and meaning of the Journal Impact Factor. *JAMA*. 2006;295(1):90-93.
112. Ramin S, Shirazi AS. Comparison between Impact factor, SCImago journal rank indicator and Eigenfactor score of nuclear medicine journals. *Nucl Med Rev*. 2012;15(2):132-136.
113. Kianifar H, Sadeghi R, Zarifmahmoudi L. Comparison between Impact Factor, Eigenfactor Metrics, and SCimago Journal Rank Indicator of pediatric neurology journals. *Acta Inform Medica*. 2014;22(2):103-106.
114. Elkins MR, Maher CG, Herbert RD, Moseley AM, Sherrington C. Correlation between the Journal Impact Factor and three other journal citation indices. *Scientometrics*. 2010;85(1):81-93.
115. Leydesdorff L. How are new citation-based journal indicators adding to the bibliometric toolbox? *J Am Soc Inf Sci Tech*. 2009;60(7):1327-1336.
116. Mahmood K, Almas K. SCImago Journal Rank Indicator: A viable alternative to Journal Impact Factor for dental journals. *LIBRES*. 2016;26(2):144-151.
117. Isaacson K. Cone beam CT and orthodontic diagnosis- A personal view. *J Orthod*. 2013;40(1):3-4.
118. Pandis N, Polychronopoulou A, Madianos P, Makou M, Eliades T. Reporting of research quality characteristics of studies published in 6 major clinical dental specialty journals. *J Evid Based Dent Pract*. 2011;11(2):75-83.
119. Schulz KF, Grimes DA. Multiplicity in randomised trials I: Endpoints and treatments. *Lancet*. 2005;365:1591-1595.
120. Feise RJ. Do multiple outcome measures require p-value adjustment? *BMC Med Res Methodol*. 2002;2:8.
121. Ranstam J. Multiple P-values and Bonferroni correction. *Osteoarthr Cartil*. 2016;24(5):763-764.
122. Zhang J, Quan H, Ng J, Stepanavage ME. Some statistical methods for multiple

endpoints in clinical trials. *Control Clin Trials*. 1997;18(3):204-221.

123. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1984;40(4):1079-1087.

124. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics*. 1987;43(3):487-498.

125. Lauter J. Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics*. 1996;52(3):964-970.

Appendices

Appendix 1 Title and abstract screening form

No.	Title	Lateral cephalometry		Cephalometric variables		Number of groups/ measurement time points for comparison	Remarks
		2D	3D	Dento-facial region	Non dento-facial region		

Appendix 2 Data extraction form

No.	Title	Year of publication (2014/2015)	Region of authorship (Americans, Europe, Asia/ Others)	Source of journal publication	Type of study design (Prospective/ retrospective)	Number of groups/ measurement time points for comparison	Number of subjects	Number & list of cephalometric variables

Level of significance- p value & adjusted p value (if any)	Each significant p value	Methods of correction	Rationale for correction	Statistician		Primary and/ or secondary outcome measures	Conclusions made according to the stated primary and/ or secondary outcome measure	Remarks
				Y	N			